2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

SINDy-CRN: Sparse Identification of Chemical Reaction Networks from Data

Nirav Bhatt[1], Bayu Jayawardhana[2], Santiago Sánchez-Escalonilla Plaza[2]

Abstract— **This work considers an important problem of identifying the dynamics of chemical reaction networks from time-series data. We propose an approach to identify complex chemical reaction networks (CRN) from concentration data using the concept of sparse model identification. Particularly, we demonstrate challenges associated with the application of the sparse identification of nonlinear dynamics (SINDy) and its variants to data obtained from CRNs. We develop a SINDy-CRN algorithm based on the properties of CRNs for identifying governing equations of a CRN. The proposed algorithm is illustrated using a numerical simulation example.**

I. INTRODUCTION

Analysis of chemical reaction networks (CRN) plays an important role in many fields, such as process industries, chemistry, systems, and synthetic biology, etc. The standard kinetic modeling of CRN involves a set of state equations that are defined based on the stoichiometry of the chemical reactions and on the reaction rate laws. A recent review on the graph-theoretical modeling framework of CRN with mass-action kinetics is presented in [1], and the extension of this to general kinetics can be found in [2]. Kinetic models of CRN have been used in literature for model-based analysis, model reduction, control, and optimization of these networks [3], [4], [5]. In biochemistry and chemical engineering, the identification of kinetic models is typically based on laborious works of isolating individual reaction and subsequently fitting the individual kinetic rate constants based on the reaction data. The obtained kinetic laws are generally given by rational functions, which makes this bottom-up approach not scalable for systems with high-dimension CRN, e.g. genome-scale kinetic modeling or complex metabolic pathways. It remains a technical challenge to identify sparse-and-yet-accurate kinetic models for these systems. Automatic identification of CRNs from concentration data is an important task in the field of CRNs and different approaches to identify governing equations from time-series data have been proposed in the existing literature [6], [7], [8], [9], [10].

*This publication is part of the project Digital Twin project 6 with project number P18-03 of the research programme Perspectief which is financed by the Dutch Research Council (NWO). Part of the work was done during a visit of Prof. Bayu Jayawardhana as a visiting faculty fellow under the Institutions of Eminence(IoE) Programme of GoI at IIT Madras, India.

[1]Nirav Bhatt is with the Department of Biotechnology, and Research Center for Data Science and AI (DSAI), Indian Institute of Technology Madras, India (email: niravbhatt@iitm.ac.in).

[2]Bayu Jayawardhana and Santiago Sánchez-Escalonilla Plaza are with the Engineering and Technology Institute Groningen, Faculty of Science and Engineering, University of Groningen, 9747AG Groningen, The Netherlands (email: b.jayawardhana@rug.nl; santiago.sanchez@rug.nl).

The identification of high-dimensional CRNs in the aforementioned complex examples typically involves identifying the underlying unknown chemical reaction networks (often called the reaction stoichiometry), and the unknown reaction-rate structures and corresponding parameters [6], [8]. Optimization problems involving all possible combinations of reaction stoichiometry candidates and reaction-rate structure are solved to identify an underlying CRN model from data [11]. Furthermore, efficient model selection methods have to be applied to discriminate different model combinations to select an appropriate model [12]. As a consequence, these approaches are computationally expensive and time-consuming. In this regard a priori information on either the reaction stoichiometry or rate structures or both proposed by human experts is often used to identify CRNs in an efficient manner [6], [13]. Often, it is difficult to obtain this CRN-specific information a priori, and it may introduce a bias in the model identification process.

Recently, sparse model identification approaches have been proposed for identifying nonlinear dynamics from time-series data to overcome the problem of model selection, i.e., selecting an appropriate model from several candidate models admitting spare solution[14], [9], [15], [16]. These approaches formulate the identification problem as a sparse optimization problem in which the dynamics of the system is a linear combination of a library of over-complete candidate functions [14], [9]. Typically, sparsity in the linear combination is induced through a regularization term in the optimization problem for obtaining a parsimonious model. Sparse identification of nonlinear dynamics (SINDy) algorithm has been proposed to discover dynamical equations from time-series data [14]. The SINDy algorithm has also been extended to handling noisy data, rational functions (implicit-SINDy), etc[17], [12], [18], [19]. Parallel and Implicit-SINDy algorithm (SINDy-PI), a variant of SINDy, can handle rational nonlinear functions and implicit function dynamics [17] and can perform efficiently model selection for noisy data. SINDy-PI has been applied to identify governing equations of biological networks [12], [20]. Although the SINDy class of algorithms can identify the underlying nonlinear dynamics from data under different conditions, the construction of an appropriate library of candidate functions is still a challenging task. Furthermore, the SINDy class of algorithms assumes that the number of state variables is equal to the number of measured variables (or concentrations of species in CRNs). However, this assumption is not valid in the case of CRNs as the number of state variables are less than the number of measured concentrations, and the

SINDy-PI cannot be applied to CRNs in a straightforward manner. Reactive SINDy proposed in [20] for identifying CRNs assumes that the reaction rates are governed by mass-action kinetics to construct a library in application of SINDy. Note that the assumption of mass-action kinetic is restrictive one and it doesn't hold for many biological and chemical engineering systems. We present a simple enzymatic example that illustrates issues that can arise in applying the SINDy-PI algorithm to CRNs that issues can not be handled by the Reactive SINDy. In this work, we present a variant of SINDy-PI method for chemical reaction networks, which we refer to as SINDy-CRN algorithm. Based on a priori knowledge of CRN properties, our proposed SINDy-CRN algorithm looks firstly at the minimal number of independent state variables from the concentration data that is closely linked to the number of independent kinetics. Subsequently, the information from the singular value decomposition of the concentration data matrix is used to construct a family of functions that can directly be used by the SINDy-PI algorithm to get a sparse representation. The efficacy of the proposed SINDy-CRN algorithm is shown via a numerical simulation example.

The paper is organized as follows. Section II describes preliminaries for SINDy-PI and models of CRNs and their properties. In Section III, the theoretical foundations and algorithm for SINDy-CRN for identifying CRNs from data is developed. Section IV demonstrates the proposed SINDy-CRN on a simulate example and Section V concludes the work.

II. Preliminaries

A. SINDy-PI

Consider the following state-space model with states $\mathbf{x} \in \mathbb{R}^n$ and measurements $\mathbf{y} \in \mathbb{R}^n$ as follows

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{f}(\mathbf{x}, \boldsymbol{\theta}) \\ \mathbf{y} &= \mathbf{x} \end{aligned} \tag{1}$$

where $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ is a vector of smooth functions, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^n$ is a a vector of measurement function, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the parameter vector. The objective of SINDy (sparse identification of nonlinear dynamics) is to identify a minimal representation $\mathbf{f}(\cdot)$ using the measurement data $\mathbf{y} = \mathbf{x}$ and $\dot{\mathbf{x}}$ for m time instances in an automated way [14][17]. It is assumed here that each element of $\mathbf{f}$ is a sparse combination of functions of $\mathbf{x}$ from a given library/kernel functions. Let $\mathbf{Y} = \mathbf{X} \in \mathbb{R}^{m \times n}$ be the measurement matrix that compiles all m measurement data for all n states, where the j-th column of the $\mathbf{X}$ matrix corresponds to the time series of the j-th state variable $\mathbf{x}_j(t_i)$, $i = 1, \dots m$. Using these notations, the SINDy problem originated from the reformulation of relationships between the measurement data $\mathbf{x}$ and $\dot{\mathbf{x}}$ as follows

$$\dot{\mathbf{X}} = \boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Sigma} \tag{2}$$

where $\boldsymbol{\Phi}(\mathbf{X}) \in \mathbb{R}^{m \times l}$ is a library matrix containing l different functions of $\mathbf{x}$ evaluated at the corresponding values of $\mathbf{X}$, and the matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{l \times n}$ is a sparse matrix containing the parameters to be identified and truncated to obtain the sparse representation of $\mathbf{f}$. Accordingly, the following regularized optimization problem can be formulated to identify $\boldsymbol{\Sigma}$

$$\min_{\boldsymbol{\Sigma}} \quad \frac{1}{2}\|\dot{\mathbf{X}} - \boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\Sigma}\|_F^2 + \alpha\gamma(\boldsymbol{\Sigma}) \tag{3}$$

where $\gamma(\boldsymbol{\Sigma})$ is a sparsity-inducing regularization function (such as l_0- or l_1- norm of $\boldsymbol{\Sigma}$), and α is a tuning parameter (hyperparameter) determining the contribution of the regularization function. Depending on the form of $\gamma(\boldsymbol{\Sigma})$, several algorithms have been proposed to solve the problem [18]. For many chemical reaction networks and biological systems, their dynamical models often involve implicit and rational dynamical functional forms of $\mathbf{x}$, where instead of (1), it takes the following implicit state-space form

$$\begin{aligned} \mathbf{f}(\mathbf{x}, \dot{\mathbf{x}}) &= \mathbf{0} \\ \mathbf{y} &= \mathbf{x}. \end{aligned} \tag{4}$$

In this case, the factorization in Eq (2) can be generalized to the implicit state-space systems as follows:

$$\boldsymbol{\Phi}(\mathbf{X}, \dot{\mathbf{X}})\tilde{\boldsymbol{\Sigma}} = \mathbf{0} \tag{5}$$

Note that the $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{l \times l}$ matrix is in the null space of $\boldsymbol{\Phi}(\mathbf{X}, \dot{\mathbf{X}})$, and hence, it is not straightforward to find the solution to the problem in Eq. (5). The SINDy-PI framework has been proposed in [17] to find a non-trivial solution of the problem in Eq. (5) through a constrained optimization formulation as follows:

$$\begin{aligned} \min_{\tilde{\boldsymbol{\Sigma}}} \quad & \frac{1}{2}\|\boldsymbol{\Phi}(\mathbf{X}, \dot{\mathbf{X}}) - \boldsymbol{\Phi}(\mathbf{X}, \dot{\mathbf{X}})\tilde{\boldsymbol{\Sigma}}\|_2^2 + \alpha\|\tilde{\boldsymbol{\Sigma}}\|_0 \\ \text{s.t.} \quad & \text{diag}(\tilde{\boldsymbol{\Sigma}}) = \mathbf{0} \end{aligned} \tag{6}$$

where $\text{diag}(\tilde{\boldsymbol{\Sigma}}) = \mathbf{0}$ denotes the diagonal elements of $\tilde{\boldsymbol{\Sigma}}$ are zero. Several convex relaxations of the optimization problem (6) have been proposed to solve the problem [17], [16]. The sequentially thresholded least-squares (STLS) is an approach to solve the problem (6) and obtain the sparse representation of $\mathbf{f}$. In the STLS, the hyperparameter λ_j, $j = 1, \dots, n$ is defined as a threshold for the jth variables. In each iteration, the parameters are pruned based on the solutions of STLS that are compared to λ_j. The algorithm will converge to a fixed set of sparse parameters that is dependent on the selection of λ_j. Sensitivity analysis w.r.t λ_j can be done to select the sparse model with the least fitting error criteria.

B. Modeling of Chemical Reaction Networks

This work considers an isothermal constant volume (V_0) chemical reaction network (CRN) involving S species and R reactions with a stoichiometric matrix, $\mathbf{N} \in \mathbb{R}^{R \times S}$, and reaction-rates vector, $\mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta}) \in \mathbb{R}^R$. Here, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the parameter vector. The reaction rate is typically a nonlinear function of the concentrations $\mathbf{c}$ and the parameter vector $\boldsymbol{\theta}$. The independent reactions can be defined as follows [21].

Definition 1 (Independent reactions): R reactions are said to be independent if (i) the rows of $\mathbf{N}$ (stoichiometries) are linearly independent, i.e., $\text{rank}\,(\mathbf{N}) = R$, and (ii) there exists some finite time interval for which the reaction rate

profiles $\mathbf{r}(t)$ are linearly independent, i.e., $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{r}(t) = 0 \Leftrightarrow \boldsymbol{\beta} = \mathbf{0}_R$.

Here, it is assumed that the R reactions are independent. Then, mole balance equations for this system can be written as

$$\begin{aligned} \dot{\mathbf{n}}(t) &= V_0 \mathbf{N}^{\mathrm{T}} \mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta}), \quad \mathbf{n}(0) = \mathbf{n}_0 \\ \mathbf{c}(t) &= \frac{\mathbf{n}(t)}{V_0}, \end{aligned} \tag{7}$$

where $\mathbf{n}$ and $\mathbf{c}$ are the S-dimensional vectors of the number of moles, and concentrations, respectively. It is assumed that the initial concentrations ($\mathbf{c}_0$) are known. The model (7) can be written in terms of the concentrations as follows:

$$\dot{\mathbf{c}}(t) = \mathbf{N}^{\mathrm{T}} \mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta}), \quad \mathbf{c}(0) = \mathbf{c}_0. \tag{8}$$

In practice, only a subset of concentrations is measured and hence, $\mathbf{c}$ be partitioned into the measured ($\mathbf{c}_m$) and unmeasured $\mathbf{c}_u$ species concentrations: $\mathbf{c}^{\mathrm{T}} = \begin{bmatrix} \mathbf{c}_m^{\mathrm{T}} & \mathbf{c}_u^{\mathrm{T}} \end{bmatrix}$. Then, the conventional state-space form for CRNs can be written as follows:

$$\begin{aligned} \dot{\mathbf{c}}(t) &= \mathbf{N}^{\mathrm{T}} \mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta}) \\ \mathbf{c}_m &= \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \end{bmatrix} \mathbf{c}. \end{aligned} \tag{9}$$

where $\mathbf{I}_m$ is the $m \times m$-dimensional matrix. The dynamic equations of $\mathbf{c}(t)$ in Eq. (9) are written in an explicit form. However, the elements of $\mathbf{r}(\mathbf{c}(t), \boldsymbol{\theta})$ for several types of CRNs can be rational functions in $\mathbf{c}$ and $\boldsymbol{\theta}$. Hence, it can be re-written in the implicit state-space form as described in Eq. (4) for the identification purpose. Hence, SINDy-PI is a suitable approach to solve this class of problems.

Following [5], a linear transformation of the concentrations in Eq. (9) can be defined as follows:

$$\begin{bmatrix} \mathbf{x}_r \\ \mathbf{x}_{iv} \end{bmatrix} = \begin{bmatrix} \mathbf{N}^{T\dagger} \\ \mathbf{Q} \end{bmatrix} (\mathbf{c} - \mathbf{c}_0), \tag{10}$$

where $\mathbf{x}_r \in \mathbb{R}^R$ is the vector of the extents of reaction states (that evolve with time), and $\mathbf{x}_{iv} \in \mathbb{R}^{S-R}$ is the vector of invariant states (that do not evolve with time). The invariant states do not change with time. Here, the symbol '†' denotes the Moore-Penrose pseudo–inverse of the matrix, and the matrix $\mathbf{Q} \in \mathbb{R}^{S-R \times S}$ is such that $\mathbf{N}\mathbf{Q}^T = \mathbf{0}$. The concentrations can be related to the reaction variant states as follows:

$$\mathbf{c}(t) = \mathbf{N}^{\mathrm{T}} \mathbf{x}_r + \mathbf{c}_0 \tag{11}$$

Correspondingly, the state-space model in form of the reaction variant (or reaction extents) states and invariants can be written as:

$$\begin{aligned} \dot{\mathbf{x}}_r(t) &= \mathbf{r}(\mathbf{x}_r(t), \boldsymbol{\theta}) \\ \dot{\mathbf{x}}_{iv} &= \mathbf{0} \\ \mathbf{c}_m &= \mathbf{N}_m^T \mathbf{x}_r + \mathbf{c}_{0,m}. \end{aligned} \tag{12}$$

The $\mathbf{c}_u$ can be computed using $\mathbf{c}_m$ and Eq. (12). We refer interested readers to [5] for the exposition of chemical reaction networks in this form and recall the following proposition from [5].

Proposition 2 ([5]): Let $\mathbf{N}$ and $\mathbf{c}$ be partitioned as: $\mathbf{N} = \begin{bmatrix} \mathbf{N}_m & \mathbf{N}_u \end{bmatrix}$ and $\mathbf{c}^{\mathrm{T}} = \begin{bmatrix} \mathbf{c}_m^{\mathrm{T}} & \mathbf{c}_u^{\mathrm{T}} \end{bmatrix}$. If $\mathrm{rank}\,(\mathbf{N}_m) = R$ then the unmeasured concentrations $\mathbf{c}_u(t)$ can be reconstructed from $\mathbf{c}_m(t)$ in two steps as follows: (i) computation of the extents of reaction, $\mathbf{x}_r(t) = (\mathbf{N}_m^{\mathrm{T}})^{\dagger}(\mathbf{c}_m(t) - \mathbf{c}_{0,m})$, and (ii) reconstruction of the unmeasured concentrations $\mathbf{c}_u(t)$: $\mathbf{c}_u(t) = \mathbf{N}_u^{\mathrm{T}}\mathbf{x}_r(t) + \mathbf{c}_{0,u}$.

The following observations can be made based on the transformation in Eq. (10) and Proposition 2.

- (O1) $\mathbf{x}_{iv}(t)$ does not change with time and depends on the initial condition, i.e., $\mathbf{x}_{iv}(t) = \mathbf{Q}(\mathbf{c} - \mathbf{c}_0) = \mathbf{Q}\mathbf{d} = \mathbf{0}_{S-R}$. The state $\mathbf{x}_{iv}(t)$ provides the $(S-R)$ relationships between $\mathbf{d}$ or $\mathbf{c} - \mathbf{c}_0$.
- (O2) The CRN in Eq. (9) can be expressed with the R differential equations describing the reaction variants.
- (O3) $\mathrm{rank}\,(\mathbf{N}_m) = R$ indicates that the minimum of R species concentrations is required to reconstruct remaining $S - R$ species concentrations. Since $S > R$, this condition is not restrictive from the identification of the CRN system.
- (O4) The unmeasured concentration $\mathbf{c}_u$ can be expressed in terms of measured concentrations: $\mathbf{c}_u(t) = \mathbf{N}_u^{\mathrm{T}}(\mathbf{N}_m^{\mathrm{T}})^{\dagger}(\mathbf{c}_m(t) - \mathbf{c}_{0,m}) + \mathbf{c}_{0,u}$.

These observations have implications for understanding the limitations of the SINDy-PI algorithm and extending the SINDy-PI to chemical reaction networks.

III. Identification of Chemical Reaction Networks using SINDy-PI

This section will extend the SINDy-PI approach to identify CRNs in a systematic manner from concentration data. The following motivation example demonstrates limitations in applying the SINDy-PI.

A. Motivating Example

Let us consider a eaction system involving a single enzymatic reversible reaction as follows: $A \rightleftharpoons B$. The mass balance equations in the concentration domain described by Eq. (9) can be written with $\mathbf{c} = \begin{bmatrix} c_a \\ c_b \end{bmatrix}$, $\mathbf{N} = [-1,\, 1]$, $\mathbf{r} = [r_1] = \frac{k_1 c_a}{k_2 + c_a} - \frac{k_3 c_b}{k_4 + c_b}$ and $\boldsymbol{\theta} = [k_1,\, k_2,\, k_3,\, k_4]^T$ where c_a and c_b are the concentrations of A and B. For numerical purposes, let the initial concentrations be given by $c_{a0} = 12$ and $c_{b0} = 3$ and let $\boldsymbol{\theta} = [5,\, 2,\, 4,\, 3]^T$. The concentrations of the components A and B are available at different time points as follows:

$$\mathbf{X} = \mathbf{C} = \begin{bmatrix} \mathbf{c}_a^T(t) \\ \mathbf{c}_b^T(t) \end{bmatrix} = \begin{bmatrix} c_a(t_1) & c_a(t_2) & \dots & c_a(t_m) \\ c_b(t_1) & c_b(t_2) & \dots & c_b(t_m) \end{bmatrix}^T$$

Then, the objective of SINDy-PI is to identify the reaction system using the concentration data $\mathbf{C}$. Let us consider the concentration measurements of $\mathbf{c}_a(t)$ to identify the differential equation of c_a variable. A library matrix of the following functional form evaluated at the different time points can be considered: $\boldsymbol{\Phi}(\mathbf{C}) = [\dot{\mathbf{c}}_a,\, \mathbf{c}_a,\, \mathbf{c}_b,\, \mathbf{c}_a\dot{\mathbf{c}}_a,\, \mathbf{c}_b\dot{\mathbf{c}}_a,\, \mathbf{c}_a\mathbf{c}_b,\, \mathbf{c}_a\mathbf{c}_b\dot{\mathbf{c}}_a]$. In order to apply the STLS algorithm, let us choose $\mathbf{y} = \Phi_j = \mathbf{c}_a\mathbf{c}_b\dot{\mathbf{c}}_a$ and the remaining term can be $\boldsymbol{\Phi}_{wj} = [\dot{\mathbf{c}}_a,\, \mathbf{c}_a,\, \mathbf{c}_b,\, \mathbf{c}_a\dot{\mathbf{c}}_a,\, \mathbf{c}_b\dot{\mathbf{c}}_a,\, \mathbf{c}_a\mathbf{c}_b]$.

Then, the solution of the least-squares $\mathbf{y} = \mathbf{\Phi}_{wj}\boldsymbol{\sigma}_j$ with $\boldsymbol{\sigma}_j$ being a coefficent vector (the jth column of $\mathbf{\Sigma}$) is $\boldsymbol{\sigma}_j = \mathbf{\Phi}_{wj}^{\dagger}\mathbf{y}$ if the $\mathbf{\Phi}_{wj}^{\dagger}$ is of the full rank $l = 6$ for $m > l$.

It can be shown that $\mathrm{rank}\,(\mathbf{\Phi}_{wj}) < l = 6$ or is a rank-deficient matrix for the reaction system. From Observation (O1), the following invariant relationship between concentrations can be established: $c_a(t) + c_b(t) = c_{a0} + c_{b0} = 15$ =constant. Then, $c_b(t) = \eta - c_a(t)$. With this relationship, the $\mathbf{\Phi}_{wj}$ can be re-written as:

$$\mathbf{\Phi}_{wj} = [\dot{\mathbf{c}}_a,\, \mathbf{c}_a,\, \mathbf{c}_b,\, \mathbf{c}_a\dot{\mathbf{c}}_a, \mathbf{c}_a\mathbf{c}_b,\, 15\dot{\mathbf{c}}_a - \mathbf{c}_a\dot{\mathbf{c}}_a]$$

By examining the $\mathbf{\Phi}_{wj}$, it can be seen that the last column can be expressed as a linear combination of the remaining columns. Hence, it is a rank-deficient matrix, and the STLS algorithm cannot be applied to this problem without re-defining the library matrix with a set of appropriate polynomial functions. With the redefined $\mathbf{\Phi}_{wj} = [\dot{\mathbf{c}}_a,\, \mathbf{c}_a,\, \mathbf{c}_b,\, \mathbf{c}_a\dot{\mathbf{c}}_a, \mathbf{c}_a\mathbf{c}_b]$, the STLS algorithm can be applied to the simulated example. The identified dynamic of the state variable c_a from the $\mathbf{c}_a$ is

$$\dot{c_a} = -\frac{15\,c_a - 8c_b + c_a c_b}{51 - c_b + c_a c_b} = -r_{1,i} \tag{13}$$

Note that the identified Eq. (13) is not of the exact form of the rate expression.

It can be shown that the simplification of the reaction rate term ($\mathbf{r} = r_1$) and the substitution of the $c_a = 15 - c_b$ in the denominator leads to the form of $r_{1,i}$. Hence, the original differential equation of c_a is identified from the data. Furthermore, the differential equation of c_b can be identified by simply differentiating the invariant form as follow: $\dot{c}_b = -\dot{c}_a = r_{1,i}$. It can be noted that the net direction of the reaction (or stoichiometric matrix corresponding to independent reaction) can also be proposed using the measurements and the identified equations as follows: $A \rightarrow B$. However, it is not possible to comment on the reversibility of the reaction without examining the identified reaction rate $r_{1,i}$ and postulating reaction kinetics.

Based on the analysis of the motivated example, the following three remarks can be made for identifying CRNs using the SINDy-PI-based approaches.

Remark 3 (Library Matrix for CRNs): The library matrix $\mathbf{\Phi}(\mathbf{X}, \dot{\mathbf{X}})$ for reaction systems may be rank-deficient due to the inherent relationships between the measured variables (concentrations) owing to the structural property of CRNs. These relationships are due to the reaction invariant states.

Remark 4 (Minimum number of state variables): The reaction networks may be represented by a less number of states than the number of measured state variables.

Remark 5 (Ambiguity in reaction rate and stoichiometry): For CRNs, it is difficult to resolve the ambiguity of individual (forward and backward reaction rates) from the identified reaction rates. The lumped reaction rate expressions can be recovered from concentration data. Furthermore, the independent reaction stoichiometry can only be recovered from concentration data. Additional information is required to resolve these ambiguities in reaction rates and stoichiometric coefficients.

Next, we will use the results from the analysis of CRN systems to extend the SINDy-PI approaches to CRN systems.

B. SINDy-PI for Chemical Reaction Network systems (SINDy-CRN)

In this section, the SIND-PI will be extended to CRN systems using the properties of CRN systems and their implications for identifying the dynamics of CRN systems from time-series data.

1) Identifying Invariant Relationships and Concentration Variables: First, we will establish the important properties of concentration data obtained from CRNs. Consider a $(m \times S)-$dimensional concentration matrix $\mathbf{C}$. The reaction variant (RV) form of concentration matrix $\mathbf{D}$ can be obtained as follows:

$$\mathbf{D} = \mathbf{C} - \mathbf{1}_m\mathbf{c}_0^T, \tag{14}$$

where $\mathbf{1}_m$ is the $m-$dimensional vector containing one as its elements. The following lemma describes the rank property of matrices $\mathbf{C}$ and $\mathbf{D}$.

Lemma 6: Consider an isothermal chemical reaction networks described by Eq. (9). Then, the concentration and the corresponding RV-form matrices can be factorized as follows

$$\mathbf{C} = \mathbf{X}\mathbf{N} + \mathbf{1}_m\mathbf{c}_0^T, \text{and } \mathbf{D} = \mathbf{X}\mathbf{N}, \tag{15}$$

and the rank of these matrices satisfies

$$\mathrm{rank}\,(\mathbf{C}) = R + 1,\ \mathrm{rank}\,(\mathbf{D}) = R \tag{16}$$

For the proof of the lemma, we refer to the same result presented in [22], [5].

Lemma 6 allows us to determine the number of reaction-variant states and invariant states for a given data matrix $\mathbf{C}$. Hence, the minimum number of state variables (Remark 4) to describe the system can be determined. The minimum number of state variables is equal to the rank of $\mathbf{D}$. Also, note that the factorization $\mathbf{D}$ with (O1) provides a hint to determine the relationship between the measured variables.

Theorem 7: Consider the concentration matrix $\mathbf{C}$ for the reaction system described by Eq. (9) with the number of observations $m \gg S$. Then, the number of independent reaction R is equal to the non-zero element of the singular value decomposition (SVD) of $\mathbf{D}^T$ and the number of invariants is $S - R$ with the basis of invariant space $\mathbf{Q}$ be given by the last $S - R$ left-singular vectors of $\mathbf{D}^T$.

Proof. Lemma 6 shows that the rank of $\mathbf{D}$ is R. The SVD of $\mathbf{D}^T$ will have R non-zero singular values and $(S - R)$ zero singular values (SV). Then,

$$SVD(\mathbf{D}^T) = \mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^T \tag{17}$$

where $\mathbf{S} = \begin{bmatrix} \mathbf{S}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_2 \end{bmatrix}$ is the $S \times S$-diagonal singular value matrix with the first $R-$ non-zero SVs (corresponding to $\mathbf{S}_1$) and $(S - R)$ zero SVs (corresponding to $\mathbf{S}_2$), and $\mathbf{U}$ and $\mathbf{V}$ are the orthogonal matrices of eigenvectors of matrices $\mathbf{D}^T\mathbf{D}$

and $\mathbf{D}\mathbf{D}^T$, respectively. By examining the $\mathbf{S}$, the number of states and the invariants relationships can be established to be R and $(S-R)$, respectively. With $\mathbf{D} = \mathbf{X}\mathbf{N}$, it can be shown that $\mathbf{Q}$ can be computed from the $\mathbf{D}$ as $\mathbf{Q}\mathbf{D}^T = \mathbf{Q}\mathbf{N}^T\mathbf{X}^T = \mathbf{0}$. Then, using Eq. (17) and the orthogonal property of the columns of $\mathbf{U}$, we get

$$\mathbf{U}_2^T\mathbf{D}^T = \underbrace{\mathbf{U}_2^T\mathbf{U}_1}_{\mathbf{0}}\mathbf{S}_1\mathbf{V}_1^T + \underbrace{\mathbf{U}_2^T\mathbf{U}_2}_{\mathbf{I}}\underbrace{\mathbf{S}_2}_{\mathbf{0}}\mathbf{V}_2^T = \mathbf{0}. \quad (18)$$

Finally, the columns of the matrix $\mathbf{U}_2^T$ provide a basis of the invariant space and $\mathbf{Q} = \mathbf{U}_2^T$. ■

Therefore, the RV-form of the concentration matrix, $\mathbf{D}$, is used to find out the number of states and the relationships between the concentration variables. Note that these relationships are linear in nature. The main implication of Theorem 7 is that only the identification of dynamical models corresponding to R species is sufficient. The dynamical models corresponding to the remaining $(S-R)$ species can be identified using the $\mathbf{Q}$ and the identified R models through algebraic manipulations. The question of the selection of the concentrations of R species for applying the SINDy-PI algorithm can be addressed by examining the $\mathbf{Q}$ matrix as given in the following corollary.

Corollary 8: Consider the identified invariant relationship $\mathbf{Q}\mathbf{d} = \mathbf{Q}(\mathbf{c} - \mathbf{c}_0) = \mathbf{0}$ for CRN as in Eq. (9). Let $\mathbf{d}$ be partitioned into independent variables $\mathbf{d}_i$ of dimension R and dependent variables $\mathbf{d}_d$ of dimension $(S-R)$ following Theorem 7, in which case the sub-matrices of $\mathbf{Q}$ corresponding to $\mathbf{d}_i$ and $\mathbf{d}_d$ are denoted by $\mathbf{Q}_i \in \mathbb{R}^{(S-R)\times R}$ and $\mathbf{Q}_d \in \mathbb{R}^{(S-R)\times(S-R)}$, respectively. Then for any partition of $\mathbf{d}$ variables satisfying $\operatorname{rank}(\mathbf{Q}_d) = S-R$, the variables $\mathbf{d}_i$ can be selected as the R species for applying the SINDy-PI algorithm.

Proof. Note that $\mathbf{Q}$ satisfies the following relationship

$$\mathbf{Q}\mathbf{d} = \mathbf{0} \Leftrightarrow \begin{bmatrix}\mathbf{Q}_d & \mathbf{Q}_i\end{bmatrix}\begin{bmatrix}\mathbf{d}_d \\ \mathbf{d}_i\end{bmatrix} = \mathbf{0} \Leftrightarrow \mathbf{Q}_d\mathbf{d}_d = -\mathbf{Q}_i\mathbf{d}_i$$

Here, $\mathbf{Q}_d$ has to be invertible to express the $(S-R)$ concentrations in terms of the R concentrations. Hence, any partition of S variables that ensures $\operatorname{rank}(\mathbf{Q}_d) = S-R$ can be used to obtain the dynamical models of $S-R$ variables from those of the R variables. In this case, $\mathbf{d}_d = \mathbf{P}\mathbf{d}_i$ with $\mathbf{P} = -\mathbf{Q}_d^{-1}\mathbf{Q}_i$. ■

Corollary 8 allows us to use only the concentration data of R species from S species for applying the SINDy-PI to identify the CRN. This addresses an important question on the data-driven identifiability of CRN based only on measurement of a subset of species concentrations. Note that several such sets of species can satisfy the rank condition. Furthermore, it can be shown that the matrix $\mathbf{P}$ is a unique matrix for a given partition.

2) Systematic Generation of Library Matrix: Corollary 8 demonstrates that once the R independent concentration variables $\mathbf{c}_i$ (or $\mathbf{d}_i$) are determined, we can apply the SINDy-PI approach using the time series of these $\mathbf{c}_i$ concentration variables to identify the dynamics of CRN. As presented before in Section II, we first need to define a suitable library matrix $\mathbf{\Phi}(\mathbf{X}, \dot{\mathbf{X}})$ which must contain the candidate terms to identify the model. The choice of the candidate terms is an important step in applying the SINDy-PI approach. Here, a priori knowledge regarding CRN can be used to select the candidate terms.

As mentioned in Remark 3, the selection of library matrix $\mathbf{\Phi}(\mathbf{X}, \dot{\mathbf{X}})$ in CRN should be done carefully to avoid rank-deficiency of $\mathbf{\Phi}(\mathbf{X}, \dot{\mathbf{X}})$ due to inherent relationships or dependencies between the variables. In Corollary 8, we have established that R independent concentration variables $\mathbf{c}_i$ can be identified solely based on data. This implies that the problem of rank deficiency due to the invariant relationships will not arise when these R independent variables are used to generate the library matrix $\mathbf{\Phi}(\mathbf{X}, \dot{\mathbf{X}})$. Based on this observation, we have the following proposition that establishes a systematic way to build a library matrix for the application of SINDy-PI.

Proposition 9: For CRN as in (9), if the R concentration variables are selected according to Corollary 8 then $\mathbf{\Phi}(\mathbf{X}, \dot{\mathbf{X}})$ is full column rank matrix.

Proof. Let us denote $\mathbf{C}_i = [\mathbf{c}_1^T, \ldots, \mathbf{c}_R^T]^T$ be the $(m \times R)-$dimensional matrix with each column representing measurements of one of the independent concentration variables. The library matrix for the R independent concentration variables $\mathbf{c}_i$ can be written as: $\mathbf{\Phi}(\mathbf{X}, \dot{\mathbf{X}}) = \mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i)$. Without loss of generality, a library matrix containing the nonlinear functions in terms of $\mathbf{c}_i$ and $\dot{\mathbf{c}}_i$ variables can be written as follows

$$\mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i) = \big[\mathbf{1}_m \quad \mathbf{C}_i \quad (\mathbf{C}_i \otimes \mathbf{C}_i) \quad (\mathbf{C}_i \otimes^2 \mathbf{C}_i) \quad \ldots \quad \ldots \quad (\mathbf{C}_i \otimes \dot{\mathbf{C}}_i) \quad \ldots\big], \quad (19)$$

where $\mathbf{a} \otimes \mathbf{b}$ denotes all unique product combinations of the components in the $\mathbf{a}$ and $\mathbf{b}$, and $\mathbf{a} \otimes^n \mathbf{b} = \mathbf{a}\underbrace{\otimes\mathbf{a} \otimes \ldots \otimes}_{n \text{ times}}\mathbf{b}$. Since $\operatorname{rank}(\mathbf{C})_i = R$, it is not possible to express any column of $\mathbf{C}_i$ in terms of other columns of $\mathbf{C}_i$. The same conclusion goes for the other terms of $\mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i)$. In this case, if the number of time instances m is greater than the size of library candidate terms, then $\mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i)$ is of full column rank. ■

As shown in the proof of Proposition 9, the library matrix $\mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i)$ can be chosen as in Eq. (19) which is guaranteed to have full column rank. This property still holds even when the library matrix has been pruned during the recursive computation of SINDy algorithm. Accordingly, the constrained optimization problem for CRN can be written as

$$\begin{aligned} \min_{\mathbf{\Sigma}} \quad & \frac{1}{2}\|\mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i) - \mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i)\mathbf{\Sigma}\|_F^2 + \alpha\|\mathbf{\Sigma}\|_0 \\ \text{s.t.} \quad & \operatorname{diag}(\mathbf{\Sigma}) = \mathbf{0}. \end{aligned} \quad (20)$$

In contrast to the optimization problem in (6), the problem in (20) uses only a subset of concentration data corresponding to the R independent concentration variables. As before the

STLS algorithm can be applied to solve the optimization problem in (20), and several sparse model candidates (one for each column σ_j of $\mathbf{\Sigma}$) will be obtained and these models can be discriminated to identify a suitable sparse model using the model selection approach described in [17].

3) SINDy-CRN algorithm: The SINDy-CRN algorithm to identify CRNs based on the previous sections' findings is presented in this section. The steps to identify CRN are presented as follows.

- **Input:** $\mathbf{C}$, $\dot{\mathbf{C}}$, thresholding parameter in the STLS, ϵ, library matrix $\mathbf{\Phi}(\mathbf{C}, \dot{\mathbf{C}})$
- **Output:** Sparse Matrix $\mathbf{\Sigma}$
- **Step 1:** Compute $\mathbf{D} = \mathbf{C} - \mathbf{1}\mathbf{c}_0^T$.
- **Step 2:** Apply SVD to $\mathbf{D}^T$; $[\mathbf{U}, \mathbf{S}, \mathbf{V}]$ =SVD($\mathbf{D}^T$).
- **Step 3:** Examine the singular values (SVs) (diagonal of $\mathbf{S}$). Assign R= Number of non-zero SVs and $S - R$ = Number of zero SVs.
- **Step 4:** Obtain $\mathbf{Q} = \mathbf{U}_2^T$ $\mathbf{U}_2$: Left singular vectors corresponding to the $S - R$ zero SVs.
- **Step 5:** Choose a valid set of independent and dependent concentration variables such that $\mathrm{rank}\,(\mathbf{Q}_d) = S - R$.
- **Step 6:** Determine the invariant relationships $\mathbf{Q}(\mathbf{c} - \mathbf{c}_0) = \mathbf{0}$.
- **Step 7:** Choose the columns of $\mathbf{C}$ corresponding to the independent concentration variables: $\mathbf{C}_i$.
- **Step 8:** Generate a library matrix $\mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i)$
- **Step 9:** Apply Sequentially thresholded least-squares (STLS) methods using $\mathbf{\Phi}(\mathbf{C}_i, \dot{\mathbf{C}}_i)$ to identify sparse matrix $\mathbf{\Sigma}$.

The columns of the identified sparse matrix $\mathbf{\Sigma}$ will be examined to discriminate different models and to obtain the R differential equations for the R independent concentration variables. Using $\mathbf{Q}$, the $(S - R)$ differential equations for the dependent concentration variables can be identified from the R differential equations.

IV. Simulation Studies

In this section, we evaluate the efficacy and validate the proposed SINDy-CRN algorithm. The following CRN is considered for the numerical simulation.

$$A + 2B \xrightarrow{r_1} C \xleftrightarrow{r_2} D. \tag{21}$$

The different matrices and vectors as in (9) are described by

$$\mathbf{c} = \begin{bmatrix} c_a & c_b & c_c & c_d \end{bmatrix}^T;\ \mathbf{N} = \begin{bmatrix} -1 & -2 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix};$$

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix};\ r_1 = \frac{k_1\, c_a\, c_b}{k_2 + k_3\, c_a\, c_b};\ r_2 = \frac{k_4\, c_c}{k_5 + c_c} - \frac{k_6\, c_d}{k_7 + c_d}$$

$$\boldsymbol{\theta} = [k_1, k_2, k_3, k_4, k_5, k_6, k_7]^T = [0.5, 3.5, 1.5, 2, 5, 1.5, 6]^T;$$

$$\mathbf{c}_0 = [1.5,\ 2.5,\ 0,\ 0]^T$$

The data matrix of concentrations $\mathbf{C} = [\mathbf{c}_a(t)\ \mathbf{c}_b(t)\ \mathbf{c}_c(t)\ \mathbf{c}_d(t)]$ is generated for the CRN (21) which will be used to validate SINDy-CRN algorithm. The SVD($\mathbf{D}^T$) leads to the following singular values: 20.77, 0.6, 0, 0. This shows that $\mathrm{rank}\,(\mathbf{D}) = 2$ and hence, the $S - R = 4 - 2 = 2$ invariant relationships can be established. The matrix $\mathbf{Q}$ corresponding to the last zero SVs is

$$\mathbf{Q} = \mathbf{U}_2^T = \begin{bmatrix} 0.1348 & 0.2697 & 0.6742 & 0.6742 \\ 0.8944 & -0.4472 & 0 & 0 \end{bmatrix}.$$

Using Step 6 in the SINDy-CRN algorithm, the two invariant relationships can be obtained after the appropriate simplification steps:

$$\begin{aligned} c_a + 2c_b + 5c_c + 5c_d &= 6.5 \\ 2c_a - c_b &= 0.5 \end{aligned} \tag{22}$$

The different partitions of the $\mathbf{Q}$ are examined and any partition with $\mathrm{rank}\,(\mathbf{Q}_d) = S - R = 2$ can be chosen. The choice of $\mathbf{Q}_d = \left[\begin{smallmatrix} 0.2697 & 0.6742 \\ -0.4472 & 0 \end{smallmatrix}\right]$ corresponding to the variables c_b and c_c satisfies the condition, and hence, they are selected as dependent variables, and the variables c_a and c_d are taken as independent variables. The corresponding time series data $\mathbf{c}_a(t)$ and $\mathbf{c}_d(t)$ will be used for identifying two differential equations by constructing a library matrix involving terms related to these independent variables.

To identify the dynamics of the independent variables, we propose two libraries which are based on the reaction (21) and the aforementioned invariant relationships (22). The first proposed library can be obtained by looking at (21) and (22). Using these relations, we construct two separate libraries as follow. For the first reaction, the library contains terms involving power of c_a up to degree 3, and its time derivative. For the second reaction depends, the library contains the terms involving c_a, c_d, their powers are up to degree 3, the mixed multiplication elements and their time derivatives. Finally, the identification mechanism used in this section is the sequentially thresholded least squares (STLS), where a parameter λ can be tuned to force a parsimonious identification of the system.

For a value of $\lambda = 0.1$ we obtained the following sparse realizations for the dynamics of the independent variables:

$$\begin{aligned} \dot{c}_a &= \frac{0.1721 \cdot c_a - 0.7188 \cdot c_a^2 + 0.2205 \cdot c_a^3}{1.51 + c_a}, \\ \dot{c}_d &= \frac{0.9299 - 0.6217 \cdot c_a - 0.9959 \cdot c_d}{1.51 + c_a \cdot c_d}. \end{aligned} \tag{23}$$

The dynamic equations of the dependent variables, c_b and c_c, can be obtained by differentiating Eq. (22) and substituting the dynamics of c_a and c_d identified in Eq. (23). Fig. 1 shows that the time evolution of the identified dynamics using the SINDy-CRN is identical to the measured concentration trajectories. However, one-to-one comparison with the simulated rates may not be possible as the multiple models can fit the concentration data.

V. Conclusions and Discussion

Identification of chemical reaction networks is an important problem in the areas of systems biology, chemistry, and chemical engineering. Developing efficient algorithms to identify CRNs from data is a still challenging task, particularly, large and complex CRNs. In this work, we proposed a SINDy-CRN algorithm to identify CRNs from

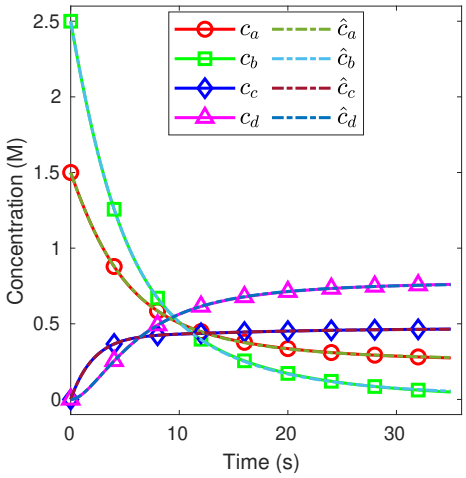

Fig. 1. Time evolution of the measured concentrations (c_a c_b, c_c, c_d) compared to the time evolution of the identified dynamics ($\hat{c}_a$ $\hat{c}_b$, $\hat{c}_c$, $\hat{c}_d$).

data by incorporating the properties of CRNs in the sparse identification of nonlinear dynamics (SINDy). First, we demonstrated using a simple example that the SINDy family of algorithms cannot be applied to data from CRNs in a straightforward manner owing to the properties of CRNs. Based on the analysis of CRNs, we demonstrated that CRNs can be modeled using a less number of state variables (minimal number of measurements) than the number of measured variables. We proposed an approach to identify the invariant relationships between state variables from the data. It is shown that the invariant relationship can be used to select appropriate concentration time-series data corresponding to the minimal number of state variables. Furthermore, the invariant relationships also help in constructing a library matrix involving candidate nonlinear functions. The SINDy-CRN algorithm was proposed by incorporating the findings of our analysis of CRN. The SINDy-CRN algorithm has finally been illustrated via a numerical example involving two enzymatic reactions.

Several open questions have to be addressed in the future for broad applications of the SINDy-CRN algorithm in the practice. This work assumes that the time-series data are noise-free and have the information content for identifying the underlying model. In the future, it is proposed to extend the SINDy-CRN to handle noisy data. In contrast to the SINDy algorithm, the SINDy-CRN uses only a subset of independent concentration time series for generating a library of candidate nonlinear functions. Hence, the resulting identified models of a CRN are functions of only the independent concentration variables. Finding physically (or biological) interpretable reaction kinetics from the identified models will be investigated in the future. The extension to the known reaction stoichiometric matrix will also be studied.

REFERENCES

[1] A. van der Schaft, S. Rao, and B. Jayawardhana, "A network dynamics approach to chemical reaction networks," *International Journal of Control*, vol. 89, no. 4, pp. 731–745, 2016.

[2] B. Jayawardhana, S. Rao, and A. van der Schaft, "Balanced chemical reaction networks governed by general kinetics," in *20th Mathematical Theory of Networks and Systems*, 2012.

[3] M. Ali Al-Radhawi, D. Angeli, and E. D. Sontag, "A computational framework for a lyapunov-enabled analysis of biochemical reaction networks," *PLoS computational biology*, vol. 16, no. 2, p. e1007681, 2020.

[4] Z. Fang, B. Jayawardhana, and A. van der Schaft, "Adaptation mechanisms in phosphorylation cycles by allosteric binding and gene autoregulation," *IEEE Trans. Automatic Control*, vol. 65, no. 8, pp. 3457–3470, 2020.

[5] N. P. Bhatt, "Extents of reaction and mass transfer in the analysis of chemical reaction systems, doctoral thesis no. 5028," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Switzerland, 2011.

[6] N. Bhatt, M. Amrhein, and D. Bonvin, "Incremental identification of reaction and mass–transfer kinetics using the concept of extents," *Industrial & engineering chemistry research*, vol. 50, no. 23, pp. 12 960–12 974, 2011.

[7] G. Russo and S. Zhuk, "On the identification of biochemical systems from intermittent and noisy data," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 3038–3043.

[8] Y. Xing, Y. Dong, C. Goergakis, Y. Zhuang, L. Zhang, J. Du, and Q. Meng, "Automatic data-driven stoichiometry identification and kinetic modeling framework for homogeneous organic reactions," *AIChE Journal*, vol. 68, no. 7, p. e17713, 2022.

[9] W. Pan, Y. Yuan, L. Ljung, J. Gonçalves, and G.-B. Stan, "Identification of nonlinear state-space systems from heterogeneous datasets," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 2, pp. 737–747, 2017.

[10] A. Papachristodoulou and B. Recht, "Determining interconnections in chemical reaction networks," in *2007 American Control Conference*. IEEE, 2007, pp. 4872–4877.

[11] A. F. Villaverde and J. R. Banga, "Reverse engineering and identification in systems biology: strategies, perspectives and challenges," *Journal of the Royal Society Interface*, vol. 11, no. 91, p. 20130505, 2014.

[12] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Inferring biological networks by sparse identification of nonlinear dynamics," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, pp. 52–63, 2016.

[13] C. J. Taylor, M. Booth, J. A. Manson, M. J. Willis, G. Clemens, B. A. Taylor, T. W. Chamberlain, and R. A. Bourne, "Rapid, automated determination of reaction models and kinetic parameters," *Chemical Engineering Journal*, vol. 413, p. 127017, 2021.

[14] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[15] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton, "Ensemble-sindy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control," *Proceedings of the Royal Society A*, vol. 478, no. 2260, p. 20210904, 2022.

[16] P. Zheng, T. Askham, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin, "A unified framework for sparse relaxed regularized regression: Sr3," *IEEE Access*, vol. 7, pp. 1404–1423, 2018.

[17] K. Kaheman, J. N. Kutz, and S. L. Brunton, "Sindy-pi: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics," *Proceedings of the Royal Society A*, vol. 476, no. 2242, p. 20200279, 2020.

[18] K. Champion, P. Zheng, A. Y. Aravkin, S. L. Brunton, and J. N. Kutz, "A unified sparse optimization framework to learn parsimonious physics-informed models from data," *IEEE Access*, vol. 8, pp. 169 259–169 271, 2020.

[19] F. Abdullah, Z. Wu, and P. D. Christofides, "Handling noisy data in sparse model identification using subsampling and co-teaching," *Computers & Chemical Engineering*, vol. 157, p. 107628, 2022.

[20] M. Hoffmann, C. Fröhner, and F. Noé, "Reactive sindy: Discovering governing reactions from concentration data," *The Journal of chemical physics*, vol. 150, no. 2, 2019.

[21] M. Amrhein, N. Bhatt, B. Srinivasan, and D. Bonvin, "Extents of reaction and flow for homogeneous reaction systems with inlet and outlet streams," *AIChE journal*, vol. 56, no. 11, pp. 2873–2886, 2010.

[22] M. Amrhein, "Reaction and flow variants/invariants for the analysis of chemical reaction data," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Switzerland, 1998.