I&EC research
Industrial & Engineering Chemistry Research

pubs.acs.org/IECR
Article

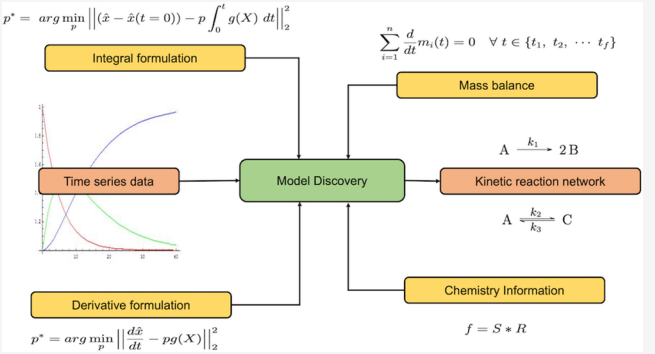
Derivative-Free Domain-Informed Data-Driven Discovery of Sparse Kinetic Models

Published as part of Industrial & Engineering Chemistry Research special issue "AI/ML in Chemical Engineering".

Siddharth Prabhu, Nick Kosir, Mayuresh V. Kothare,* and Srinivas Rangarajan*

Cite This: *Ind. Eng. Chem. Res.* 2025, 64, 2601−2615

ABSTRACT: Developing data-driven kinetic models from reaction data is valuable for inferring the underlying reactions and designing reactive processes without needing first-principles models. However, recently developed techniques to learn interpretable dynamical models from data are susceptible to inherent experimental noise, especially in reaction kinetics data. Here, we address these issues by (1) employing a new derivative-free technique for sparse identification of dynamical equations that approximates the integral rather than the derivative (which we call as *DF-SINDy*) and (2) including domain information such as mass balance and chemistry information. We demonstrate this using retrospective examples to recover the true (known) governing equations from synthetic data under varying noise levels, sampling frequencies, and number of experiments. We observe that (1) models discovered from *DF-SINDy* have lower errors than those discovered from *SINDy* (*Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 3932−3937, DOI: 10.1073/pnas.1517384113) and (2) adding domain knowledge further helps recover correct terms, thereby improving the reliability of the interpretations obtained from these models. This work is chemistry agnostic and represents a step toward developing domain-informed interpretable kinetic models for complex reaction networks.



INTRODUCTION

Kinetic models of catalytic reactions allow for (1) predicting rates and yields at new conditions to ultimately design new processes and (2) infer the underlying mechanistic details of the system. The state-of-the-art mechanistic kinetic models are developed using density functional theory (DFT) based microkinetic modeling along with infusion of experimental kinetics data.[2−5] However, for complex catalytic materials that comprise multiple functionalities, dynamically restructuring surfaces, large reaction networks, etc., such an approach can still be computationally intractable requiring data-driven model surrogates to reduce the computational burden.[6−12] In these cases, learning kinetic models directly from the data can be a valuable starting point.[6,13] Black-box neural networks,[14] kinetic-informed neural networks,[15,16] and Bayesian equation discovery[17] are some examples of data-driven kinetic models. However, we posit that learning *interpretable* models directly from experimental kinetic data would advance our understanding of the underlying reaction mechanism, which could then inform subsequent experiments or ab initio studies. Advances in machine learning theories and algorithms and data acquisition techniques have encouraged the development of methods to discern governing laws or constitutive equations directly from data.[18−28]

An emerging technique for developing interpretable kinetic models is sparse identification of nonlinear dynamics (*SINDy*).[1,29,30] This method generates differential equation models of observable states that contain a few dominant terms from a candidate library of user-specified choices, resulting in models that are sparse and, consequently, interpretable. *SINDy* models can be built using only measurable data, and therefore, the states retain their physical interpretation. The terms identified by *SINDy* can, in principle, be explained to derive mechanistic insights. For instance, if *SINDy* identifies a term such as xy in $\frac{dx}{dt}$, it suggests the presence of a reaction involving x and y. While the exact sequence of elementary steps remains unknown, this provides valuable insights into the underlying

Received: August 9, 2024
Revised: January 11, 2025
Accepted: January 13, 2025
Published: January 27, 2025

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

processes. Furthermore, the learning process primarily involves linear regression (with L_2 penalties), therefore, *SINDy* models are straightforward to train. As a result, *SINDy* models have been used to discover partial differential equations[31] and ordinary differential equations in biological systems,[30] chaotic systems,[1] aeronautical systems,[32] etc. Its use in building reaction models has been attempted but has remained relatively less explored.[29,33,34]

Reaction experiments are often noisy; they do not close mass or atom balances due to instrument calibration errors, incomplete characterization of products, side reactions leading to coke, etc. Further, instrumentation or experimental design constraints may lead to insufficient sampling of the reaction system, thereby missing important transients in concentrations. In such contexts, a limitation of the original *SINDy* formulation is that since derivative values are required at each time point, noisy or sparsely sampled data can compromise the accuracy of the discovered models and their generalizability to unseen data. Previous work addresses these challenges in three broad ways. The first approach involves denoising the measurements[35] or their derivatives[36,37] and subsequently learning a model using methods such as *SINDy*. The two steps, denoising and learning, can also be combined into a single optimization problem.[38] This approach may lead to solving either an overall nonconvex problem or approximating the derivatives, which can lead to errors when data are not sampled frequently enough. The second approach avoids approximating derivatives of measurements and instead trains using the measured data alone. This requires solving the assumed differential equation, which is usually accomplished using collocation techniques[39,40] although numerical integration techniques may also be employed. This approach can account for noisy and sparse measurements with large sampling times and can even include unmeasured states in the model; however, the optimization problem that is solved is always nonconvex and, therefore, is computationally costlier than *SINDy* and is guaranteed to converge only to a local optimum. The third approach directly learns from noisy measurements either using ensemble or Bayesian techniques.[41−44] This approach can be computationally demanding when bootstrapping or subset selection is used, especially with a large function library. If derivatives are computed via finite difference, they can still compromise the accuracy of the model.

In this work, we address the challenge of noisy data and generalizability in the context of kinetic models in two ways. First, we employ a derivative-free *SINDy*-like approach, which we term *DF-SINDy*, that is inspired by the formulation of the integral problem.[45,46] This approach eliminates the need to approximate the derivatives of the measurements, keeps the optimization problem convex in most cases, and retains the flexibility and interpretability of *SINDy*. Second, we include physics- and chemistry-based constraints, so that the inferred models do not violate known domain knowledge. We hypothesize that such constraints will prevent *DF-SINDy* from proposing meaningless terms that may otherwise be identified in the presence of noisy or limited data.

The contents of the article are as follows. First, we describe the unconstrained formulation of *DF-SINDy*. Second, we propose two domain-informed versions, namely, a mass balance formulation and a chemistry-informed formulation. Third, we test these formulations using retrospective examples of recovering the true governing equations of a known reaction model under increasingly *imperfect* data conditions, i.e., at increased noise levels, fewer experiments, and reduced sampling frequency. We show that the three *DF-SINDy* formulations outperform traditional *SINDy* when the data are imperfect. Additionally, to mimic realistic temperature controls, we consider the problem of learning the Arrhenius-type temperature dependence of kinetic parameters from isothermal experiments conducted at multiple operating temperatures. We also consider a stiff example to show the importance of sampling in identifying the equations. We show that a *DF-SINDy* like loss function (based only on the states) converges to a better local minimum than the *SINDy* like loss function (based on states and derivatives). Finally, we consider a case of overestimating a reaction mechanism by adding extra reactions and assuming an irreversible reaction to be reversible. We show that our method can identify the correct terms in the model.

METHOD

Reaction kinetics experiments typically involve (1) measuring the progress of a reaction system in batch experiments by collecting concentration−time data for stable species (reactants, intermediates, and products) for varying initial conditions and temperature or (2) measuring exit concentrations of stable species in flow reactors for different residence times, inlet conditions, and temperatures. Our goal is to utilize such data (since time and residence time can be used interchangeably for a batch/plug flow reactor, we only consider batch systems here) to derive a kinetic model, i.e., a set of ordinary differential equations (ODE) of the form shown in eq 1, which holds for any reaction system (regardless of the phase).

$$\frac{\mathrm{d}}{\mathrm{d}t}X(t) = f(X, T, p) \tag{1}$$

The vector $X(t) = [x_1(t), x_2(t), ..., x_n(t)]^T \in \mathbb{R}^{n\times 1}$ represents the concentration of measured species in the system (n in total) at time t. T represents the temperature, which can either be constant for a reaction or change with respect to time, and p is the parameter of the differential equations. The functions $f: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ describe the dynamics of the system that we wish to learn from the measurements.

Unconstrained Formulation. Our initial (or naive) formulation is one where we assume that the temperature is fixed in all of the experiments. Each experiment is carried out with differing initial concentration $X(t = 0)$ and sampled periodically to collect a data set $\hat{X}(t)$. For this case, the dynamics are expected to be of the form given in eq 2.

$$\frac{\mathrm{d}}{\mathrm{d}t}x_k(t) = f_k(X(t), p_k) = g_k(X(t))\cdot p_k \quad \forall\ k \in \{1, 2, \cdots n\} \tag{2}$$

We assume that the parameters p_k appear linearly in f_k (which is reasonable given that we can assume mass action kinetics) and are constant; that is, they do not vary with time. f_k can therefore be decomposed as a dot product $(\cdot)$ between features $g_k(X)$ and parameters p_k. Then, the solution to eq 2 is given by

$$\begin{aligned} x_k(t) &= x_k(t = 0) + \int_0^t g_k(X(t))\cdot p_k \,\mathrm{d}t \\ &= x_k(t = 0) + p_k \cdot \int_0^t g_k(X(t)) \,\mathrm{d}t \quad \forall\ k \in \{1, 2, ..., n\} \end{aligned} \tag{3}$$

To find the parameters p_k, one can formulate a least-squares problem, dropping the subscript k and the explicit notation of time dependence of $X(t)$ for convenience, as follows (eq 4).

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

$$p^* = \arg\min_p \|(\hat{x} - \hat{x}(t=0)) - p\int_0^t g(X)\,dt\|_2^2 \quad (4)$$

Because parameters p appear linearly, the problem is convex and therefore its analytical solution can be written as eq 5.

$$p^* = \left[\left(\int_0^t g(X)\,dt\right)^T\left(\int_0^t g(X)\,dt\right)\right]^{-1}\left(\int_0^t g(X)\,dt\right)^T (\hat{x}(t) - \hat{x}(t=0)) \quad (5)$$

Since we do not know $g(X)$ a priori, we consider a functional library vector $\Theta(X)$ of b terms that contains all possible polynomial combinations of concentrations of all species (as reaction kinetics models usually contain polynomial terms) as shown in eq 6 and assume that $g(X) \in \Theta(X)$.

$$\Theta(X) = [x_1 \quad x_2 \ldots x_n \quad x_1^2 \quad x_1x_2 \ldots x_n^2 \quad x_1^3 \quad x_1x_2x_n \ldots] \quad (6)$$

These polynomial terms are user-defined and are akin to the original *SINDy* method. Since $g(X) \in \Theta(X)$, to calculate $\int_0^t g(X)\,dt$, we need $x_k \in X$ as an explicit function of time. We approximate this explicit function of all the concentrations $x_k \in X$ using cubic spline interpolation[47] on the measurements $\hat{x}_k$. Let $\Psi(\hat{X}) = \{\psi_1(\hat{x}_1)\psi_2(\hat{x}_2)\ldots\psi_n(\hat{x}_n)\}$ be a set of interpolating functions obtained from measurements $\hat{X}$, then integrating the library matrix results in eq 7.

$$\int_{t_1}^{t_2}\Theta(X)\,dt \approx \int_{t_1}^{t_2}\Theta(\Psi(\hat{X}))\,dt = \left[\int_{t_1}^{t_2}\psi_1\,dt \quad \int_{t_1}^{t_2}\psi_2\,dt \ldots \int_{t_1}^{t_2}\psi_n\,dt \quad \int_{t_1}^{t_2}\psi_1^2\,dt \quad \int_{t_1}^{t_2}\psi_1\psi_2\,dt \ldots \int_{t_1}^{t_2}\psi_n^2\,dt \quad \int_{t_1}^{t_2}\psi_1^3\,dt \ldots\right] \quad (7)$$

Consider the coefficient matrix $\Xi = [\xi_1, \xi_2, \ldots, \xi_n] \in \mathbb{R}^{b\times n}$ where each column ξ_i represents the coefficients corresponding to the terms in the functional library matrix $\Theta(\Psi(\hat{X}))$, then the linear system identification problem is formulated as shown in eq 8. We also introduce a multiplicative matrix $M \in \mathbb{R}^{n\times n}$ which is an identity matrix here but will allow incorporating mass balance and chemistry information in future sections.

$$\begin{bmatrix} \hat{x}_1(t) - \hat{x}_1(t_0) \\ \hat{x}_2(t) - \hat{x}_2(t_0) \\ \vdots \\ \hat{x}_{n-1}(t) - \hat{x}_{n-1}(t_0) \\ \hat{x}_n(t) - \hat{x}_n(t_0) \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{n\times n} \begin{bmatrix} \int_0^t \Theta(\Psi)\,dt\cdot\xi_1 \\ \vdots \\ \int_0^t \Theta(\Psi)\,dt\cdot\xi_n \end{bmatrix}_{n\times 1} \quad (8)$$

To find the optimal value of the coefficient matrix Ξ, the optimization problem in eq 4 can be written as shown in eq 9, where the summation is over all the measurements obtained between the start time $t = t_0$ and the end time $t = t_f$. In this work, we refer to this formulation as the unconstrained formulation. The number of decision variables in this optimization problem (eq 9) is bn

$$\Xi^* = \arg\min_{\Xi}\frac{1}{2}\sum_{t=t_1}^{t_f}\text{MSE (Equation 8)} \quad (9)$$

Algorithm. Since $g(X) \in \Theta(X)$, we need to find a sparse solution to the optimization problem given in eq 9. We use sequential threshold least-squares *(STLSQ)*[1] algorithm that minimizes a least-squares objective function with ridge penalty (λ). It sets the coefficients that are less than the thresholding parameter (ϵ) to zero and solves the optimization problem again. Note that other variable selection algorithms such as LASSO, Elastic Net, SR3, etc., can also be used. The modified optimization problem is given in eq 10.

$$\Xi^* = \arg\min_{\Xi}\frac{1}{2}\left[\sum_{t=t_1}^{t_f}\text{MSE (Equation 8)} + \lambda\sum_{i=1}^{n}\|\xi_i\|_2^2\right] \quad (10)$$

There are two possible terminating conditions: (1) All the coefficients in matrix Ξ are eliminated, in which case the thresholding is too large to consider all the coefficients and no solution is obtained. (2) There are no more coefficients to be eliminated, in which case the optimal solution with the remaining coefficients has been found. If the algorithm eliminated all of the coefficients, then consider either lowering the thresholding parameter or changing the terms in the polynomial library. Once the algorithm is terminated successfully, the values of the coefficients are returned. The steps are outlined in algorithm 2. We use an interior point solver called IPOPT[48] from CasADi[49] to solve the optimization problem. We also perform hyperparameter optimization given in algorithm 1 to select the best model based on eq 11. The various hyperparameters used are given in the Supporting Information (SI) (Tables S1−S6).

$$\text{metric} = 2\cdot\log\left(\frac{MSE}{2}\right) + \text{complexity} \quad (11)$$

Algorithm 1 Hyperparameter Optimization

Require: Hyperparameters : ridge penalty λ, thresholding parameter ϵ and polynomial terms in functional library matrix
$H \leftarrow$ set ▷ Initialize set of all possible combinations of hyperparameters.
score $\leftarrow$ list ▷ Initialize empty list
for $h_i \in H$ **do** in parallel
  model $\leftarrow$ Algorithm 2 ▷ Discover models
  score $\leftarrow$ Append(Equation 11) ▷ Calculate the metric for the discovered model
end for
BestModel $\leftarrow$ min(score) ▷ Select the model with the minimum metric value
return BestModel

Algorithm 2 Sequential Threshold Least Square

Require: Measurement matrix $\dot{X}$, candidate library $\Theta(\Psi)$, multiplicative matrix M, ridge penalty λ, thresholding parameter ϵ
$\Xi_{guess} \sim \mathcal{U}(\epsilon, \infty)$ ▷ Initialize from a uniform distribution
$I_{big} \leftarrow \Xi \geq \epsilon$ ▷ Initialize boolean matrix
while not converged **do**
  $\Xi^* \leftarrow$ Solve the optimization problem 10 or 15 or 18 with $\Xi(I_{big})$ as decision variables
  $I_{small} \leftarrow \Psi \leq \epsilon$ ▷ Find small indices
  $\Xi(I_{small}) \leftarrow 0$ ▷ Threshold small indices to zero
  $I_{big} \leftarrow\sim I_{small}$ ▷ Update the boolean matrix
  $\Xi_{guess} \leftarrow \Xi^*$ ▷ Initialize with optimal values of previous iteration
end while
return Ξ^*

Mass Balance Formulation. In this formulation, we use the fact that in a batch system, the sum of the masses of all the species at any point is equal to the sum of the masses of those species at the initial point, i.e., mass is always conserved.

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

Therefore, the derivative of the sum of the masses is zero, however the individual masses may change as described in eq 12.

$$\sum_{i=1}^{n} m_i(t) = \sum_{i=1}^{n} m_i(t_0) \quad \forall\, t \in \{t_1, t_2, \ldots t_f\}$$
$$\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}t} m_i(t) = 0 \quad \forall\, t \in \{t_1, t_2, \ldots t_f\} \tag{12}$$

Multiplying and dividing eq 12 by the molecular weight (w_i) of the corresponding species gives the rate of change of moles.

$$\sum_{i=1}^{n} \frac{w_i}{w_i} \frac{\mathrm{d}}{\mathrm{d}t} m_i(t) = 0$$
$$\text{or} \sum_{i=1}^{n} w_i \frac{\mathrm{d}}{\mathrm{d}t} x_i(t) = 0 \quad \text{since } x_i \propto \frac{m_i}{w_i} \text{ for a fixed volume}$$
$$\text{i.e.} \sum_{i=1, i\neq k}^{n} \frac{w_i}{w_k} \frac{\mathrm{d}}{\mathrm{d}t} x_i(t) = -\frac{\mathrm{d}}{\mathrm{d}t} x_k(t) \quad \forall\, t \in \{t_1, t_2, \ldots t_f\} \tag{13}$$

For $k = n$, eq 13 can be written in a matrix form where the n_{th} row in the multiplicative matrix M in eq 8 is replaced by the ratio of molecular weights of individual components by the n_{th} component as shown in eq 14

$$\begin{bmatrix} \hat{x}_1(t) - \hat{x}_1(t_0) \\ \hat{x}_2(t) - \hat{x}_2(t_0) \\ \vdots \\ \hat{x}_{n-1}(t) - \hat{x}_{n-1}(t_0) \\ \hat{x}_n(t) - \hat{x}_n(t_0) \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\frac{w_1}{w_n} & -\frac{w_2}{w_n} & \cdots & -\frac{w_{n-1}}{w_n} \end{bmatrix} \begin{bmatrix} \int_0^t \Theta(\Psi)\, \mathrm{d}t \cdot \xi_1 \\ \vdots \\ \int_0^t \Theta(\Psi)\, \mathrm{d}t \cdot \xi_{n-1} \end{bmatrix} \tag{14}$$

Consequently, the optimization problem with $(n-1)b$ decision variables can be written as follows (eq 15), where the first term of the loss is the mean squared error of the measured states and model predictions (using eq 14) for different times.

$$\Xi^* = \arg\min_{\Xi} \frac{1}{2} \left[\sum_{t=t_1}^{t_f} \text{MSE (Equation 14)} + \lambda \sum_{i=1}^{n-1} \|\xi_i\|_2^2 \right] \tag{15}$$

Chemistry Formulation. A plausible reaction network is often known to the expert (or can be postulated);[39] in such a case, we can say that

$$f = S\dot{R} \tag{16}$$

where in eq 16, $S \in \mathbb{R}^{n\times r}$ is the stoichiometric matrix of the network (n represents the number of reactants and products and r represents the number of reactions in the system) is a stoichiometric matrix of the reaction network and $R \in \mathbb{R}^{r\times 1}$ is the rate of these reactions. In such a scenario, we can replace the multiplicative matrix M by the stoichiometric matrix S and write $R = \Theta(X)\xi$. Further, often, the rates depend only on the reactants and product concentrations. Therefore, $\Theta(X)$ is no longer a polynomial combination of all the species but instead a polynomial combination of only those species that the particular reaction depends on. Consequently $\Theta(X)$ can be denoted as $\Theta^{(r)}(X)$ and the coefficients as $\Xi \in \{\xi_1, \xi_2, \ldots \xi_r\}$ with the vector ξ_r as the coefficients corresponding to the columns of $\Theta^{(r)}(X)$ since they are now reaction dependent. It should be noted that the stoichiometric matrix of balanced reactions naturally enforces mass balance. Additionally, this approach differs from parameter estimation of a given reaction mechanism in that, with parameter estimation, the analytical equations for the rates of consumption/formation are known and only the kinetic or thermodynamic constants are estimated, while in this approach, we learn both the rate expressions and the parameters.

$$\begin{bmatrix} \hat{x}_1(t) - \hat{x}_0(t_0) \\ \hat{x}_2(t) - \hat{x}_0(t_0) \\ \vdots \\ \hat{x}_{n-1}(t) - \hat{x}_{n-1}(t_0) \\ \hat{x}_n(t) - \hat{x}_n(t_0) \end{bmatrix} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,r} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,r} \\ \vdots & \vdots & & \vdots \\ s_{n-1,1} & s_{n-1,2} & \cdots & s_{n-1,r} \\ s_{n,1} & s_{n,2} & \cdots & s_{n,r} \end{bmatrix} \begin{bmatrix} \int_0^t \Theta^{(1)}(\Psi)\, \mathrm{d}t \cdot \xi_1 \\ \vdots \\ \int_0^t \Theta^{(r)}(\Psi)\, \mathrm{d}t \cdot \xi_r \end{bmatrix} \tag{17}$$

Consequently the optimization problem with rc decision variables (such that $r \leq n$, $c \leq b$) is written as follows (eq 18), wherein the first term of the loss is a mean squared error between the measurements and the model prediction (using eq 17) over all sampled times.

$$\Xi^* = \arg\min_{\Xi} \frac{1}{2} \left[\sum_{t=t_1}^{t_f} \text{MSE (Equation 17)} + \lambda \sum_{i=1}^{r} \|\xi_i\|_2^2 \right] \tag{18}$$

Nonlinear Least-Squares Problem: Isothermal Process Conditions. The reaction systems given by eq 2 assume constant parameters that appear linearly in the model. However, when reactions are conducted at different temperatures, reaction rates (therefore, the parameters) depend on the temperature. If the rate constants follow the Arrhenius equation, then the dynamical equation has the form given in eq 19.

$$\frac{\mathrm{d}}{\mathrm{d}t} x_k = f_k(X, p_k) = g_k(X) \cdot h(p_k, T) =$$
$$g_k(X) \cdot (k(T = T_{ref}) \otimes \mathrm{e}^{-E/R(\frac{1}{T} - \frac{1}{T_{ref}})}) \tag{19}$$

where parameters p_k are as follows: $k(T = T_{ref})$ is a vector of reaction rate constants at any given reference temperature (T_{ref}), E is a vector of activation energies, R is the universal gas constant, and $\otimes$ is the element-wise product. When concentration values are measured from isothermal experiments, i.e., the temperature is constant in an experiment but can vary from one to the next, $h(p, T)$ is independent of time. Thus, one can formulate a nonlinear least-squares problem in parameters given by eq 20.

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

Table 1. Simplified Reaction Mechanism of Cracking and Isomerization of Butenes

R_i	Reaction	Forward reaction	Reverse reaction
1	A → 2B	$k_1 = 4.283$, $E_1 = 30000$	
2	A ↔ C	$k_2 = 1.191$, $E_2 = 40000$	$k_3 = 5.743$, $E_3 = 45000$
3	A ↔ D	$k_4 = 10.219$, $E_4 = 50000$	$k_5 = 1.535$, $E_5 = 60000$

Table 2. Reaction Mechanism of Esterification of Carboxylic Acid

R_i	Reaction	Forward reaction	Reverse reaction
1	F + D ↔ E + X	$k_1 = 0.3$	$k_7 = 1.2$
2	E + B ↔ G + D	$k_2 = 0.4$	$k_8 = 0.6$
3	A + G ↔ E + C	$k_3 = 1.1$	$k_9 = 0.7$
4	B + D ↔ J + E	$k_4 = 0.9$	$k_{10} = 0.1$
5	A + E ↔ D + K	$k_5 = 1.0$	$k_{11} = 0.5$
6	X + K ↔ L + D	$k_6 = 0.2$	$k_{12} = 0.8$

Table 3. List of Parameters of the System for Three Different Training Conditions

RN_i	System parameters	Noise	Experiments	Sampling rate
1	Initial conditions	$\mathcal{U}(5, 20)$	$\mathcal{U}(5, 20)$	$\mathcal{U}(5, 20)$
	Number of experiments (nexpt)	6	2, 4, 6	6
	sampling rate (Δt)	0.01	0.05	0.01, 0.05, 0.1
	standard deviation of noise (σ)	0, 0.1, 0.2	0	0
2	Initial conditions	$\mathcal{U}(5, 20)$	$\mathcal{U}(5, 20)$	$\mathcal{U}(5, 20)$
	Number of experiments (nexpt)	25	15, 20, 25	25
	sampling rate (Δt)	0.01	0.01	0.01, 0.05, 0.1
	standard deviation of noise (σ)	0, 0.1, 0.2	0	0

$$p^* = \arg\min_p \|(\hat{x}(t) - \hat{x}(t=0)) - h(p, T)\cdot\int_0^t g(X)\,\mathrm{d}t\|_2^2 \tag{20}$$

The main difference between this formulation and the previous ones is that each element in Ξ is written as eq 21.

$$\xi_i = \xi_i^{(1)}(T = T_{ref}) \otimes \mathrm{e}^{-\xi_i^{(2)}/R(\frac{1}{T}-\frac{1}{T_{ref}})} \tag{21}$$

where $\xi_i^{(1)}(T = T_{ref})$ is the reaction rate at the reference temperature and $\xi_i^{(2)}$ is the activation energy.

Unlike the linear least-squares formulations previously discussed, this is a nonconvex problem and therefore does not have an analytical solution. However, we note that the integration in eq 20 is straightforward and can be solved using numerical techniques as a system of uncoupled differential equations.

Kinetic Reaction Network. *RN1: Cracking and Isomerization of Butene.* We consider a simplified network of cracking and isomerization of butenes,[50] a representative chemistry for breaking down large hydrocarbons to smaller olefins. We assume that the cracking reaction is irreversible while the isomerization reactions are reversible and that all steps are pseudo first order. The overall reaction network is given in Table 1, and the reaction rates are in eq 22. The forward/reverse reaction rate constant is denoted by k, the activation energy in J/mol is denoted by E, and A, B, C, D correspond to 1-butene, ethene, 2-butene, isobutene, respectively. The kinetic parameters have been scaled down from the original source to remove multi-time-scale behavior and reduce stiffness; further, reasonable activation barrier values were assumed for the steps.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}A\\B\\C\\D\end{bmatrix} = \begin{bmatrix}r_0\\r_1\\r_2\\r_3\end{bmatrix} = \begin{bmatrix}-1&-1&-1\\2&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}R_1\\R_2\\R_3\end{bmatrix} = \begin{bmatrix}-1&-1&-1\\2&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}k_1A\\k_2A - k_3C\\k_4A - k_5D\end{bmatrix} = \begin{bmatrix}-15.693A + 5.743C + 1.534D\\8.566A\\1.191A - 5.743C\\10.219A - 1.535D\end{bmatrix} \tag{22}$$

When the reaction rates depend on temperature, the dynamic equation is given as eq 23.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}A\\B\\C\\D\end{bmatrix} = \begin{bmatrix}-1&-1&-1\\2&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}k_1\mathrm{e}^{-E_1/R(\frac{1}{T}-\frac{1}{373})}A\\k_2\mathrm{e}^{-E_2/R(\frac{1}{T}-\frac{1}{373})}A - k_3e^{-E_3/R(\frac{1}{T}-\frac{1}{373})}C\\k_4\mathrm{e}^{-E_4/R(\frac{1}{T}-\frac{1}{373})}A - k_5\mathrm{e}^{-E_5/R(\frac{1}{T}-\frac{1}{373})}D\end{bmatrix} \tag{23}$$

RN2: Esterification of Carboxylic Acid. We also consider a more complex reaction network of esterification of carboxylic acid.[51] The overall reaction network is summarized in Table 2. This reaction network consists of six reversible reactions and 11 different species, the reaction rates of which are given in eq 24.

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

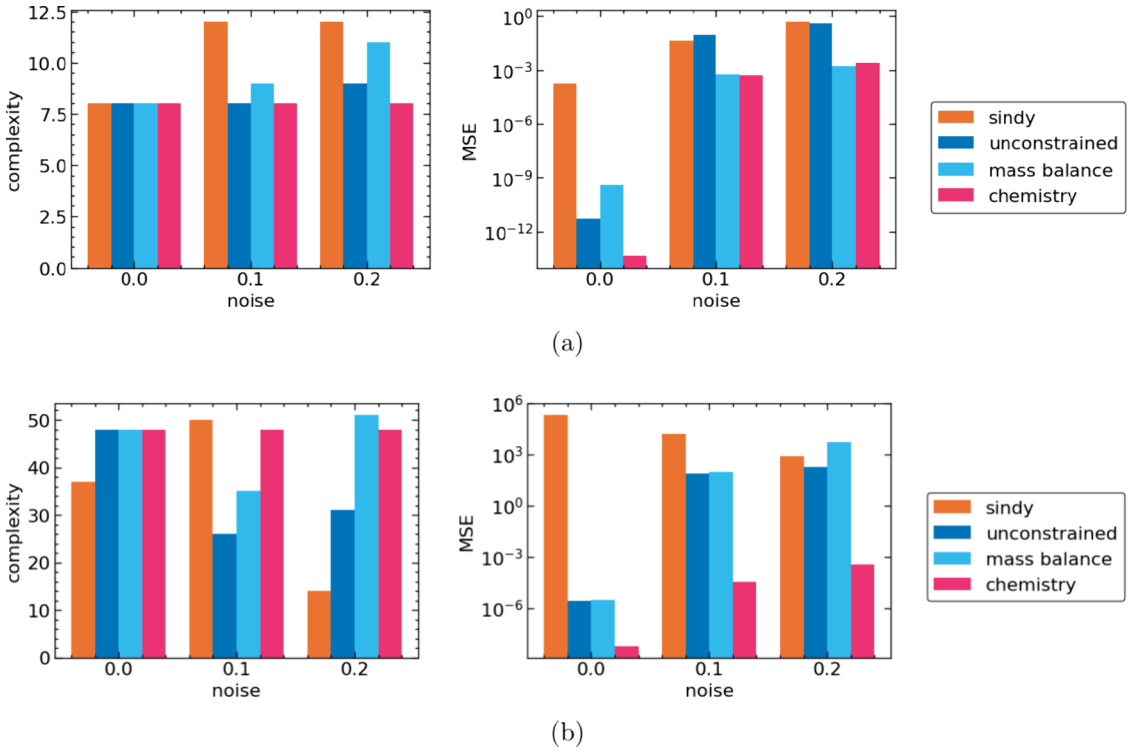


Figure 1. A comparison of model complexity and mean square testing error (MSE) for RN1 (a) and RN2 (b) is shown for the three linear *DF-SINDy* formulations (unconstrained, mass balance, and chemistry) compared with the naive *SINDy* for different levels of noise (i.e., increasing standard deviation of Gaussian noise).

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} A \\ B \\ C \\ D \\ E \\ F \\ G \\ J \\ K \\ L \\ X \end{bmatrix} = \begin{bmatrix} 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} k_1FD - k_7EX \\ k_2EB - k_8GD \\ k_3AG - k_9EC \\ k_4BD - k_{10}JE \\ k_5AE - k_{11}DK \\ k_6XK - k_{12}LD \end{bmatrix} \tag{24}$$

Generation of Synthetic Data. To synthesize testing data, we integrate the dynamics using *odeint* in python with random initial conditions uniformly distributed in [5, 10]. We record data up to 10 s with a time step of 0.01 s and 6 different initial conditions (Table 3). We use the same testing data for all cases and approaches to rigorously compare the performance of the discovered models across different training conditions. Training data depends on the three training conditions and is summarized in the following table. We note that while a sampling rate of 0.01 s is unlikely with methods such as chromatography, such rates can be feasible with techniques such as infrared spectroscopy or mass spectrometry.

For training the nonlinear least-squares problem, we consider the same set of parameters except the temperature, which now changes in every experiment and is chosen between 360 K and 385 K.

RESULTS AND DISCUSSION

Increasing Noise Levels. We first study the effect of noise on the recovery of the true model from the unconstrained, mass balance, and chemistry-based formulations with naive *SINDy* as our reference. We add Gaussian noise with two different standard deviation values (0.1 and 0.2) to the synthesized training data while being mindful that adding excessive noise can lead to irrecoverable conditions.[52] We then perform hyperparameter optimization and select the model that minimizes eq 11. Once the model is chosen, we report the complexity (number of retained terms) and mean squared errors on the test set. Figure 1a for RN1 and Figure 1b for RN2 provide a comparison of these approaches for model complexity and mean squared errors at varying noise levels. In the absence of noise, all methods recover the true complexity for RN1. However, for RN2 without noise, *SINDy* fails to recover the correct terms while all other formulations accurately identify them. This suggests that our integral formulation requires fewer data compared to *SINDy* to discover accurate models. As the noise level increases (i.e., at higher standard deviations of Gaussian noise), *SINDy* starts to retain additional terms that are not in the original model (for RN1) or drop terms that are in the original model (especially for RN2). The unconstrained and mass balance *DF-SINDy* formulations also produce models with incorrect complexity as *SINDy*, however, their MSE is typically 1−2 orders of magnitude lower, emphasizing the relative accuracy of loss functions that are based on state measurements rather than their derivatives (as in the case of *SINDy*). In particular, finite differences of noisy data further amplify the noise in measurements and thus lead to poorer recovery. The chemistry formulation provides the best recovery; it is robust to noise, and the MSE is considerably lower than the other three formulations.

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

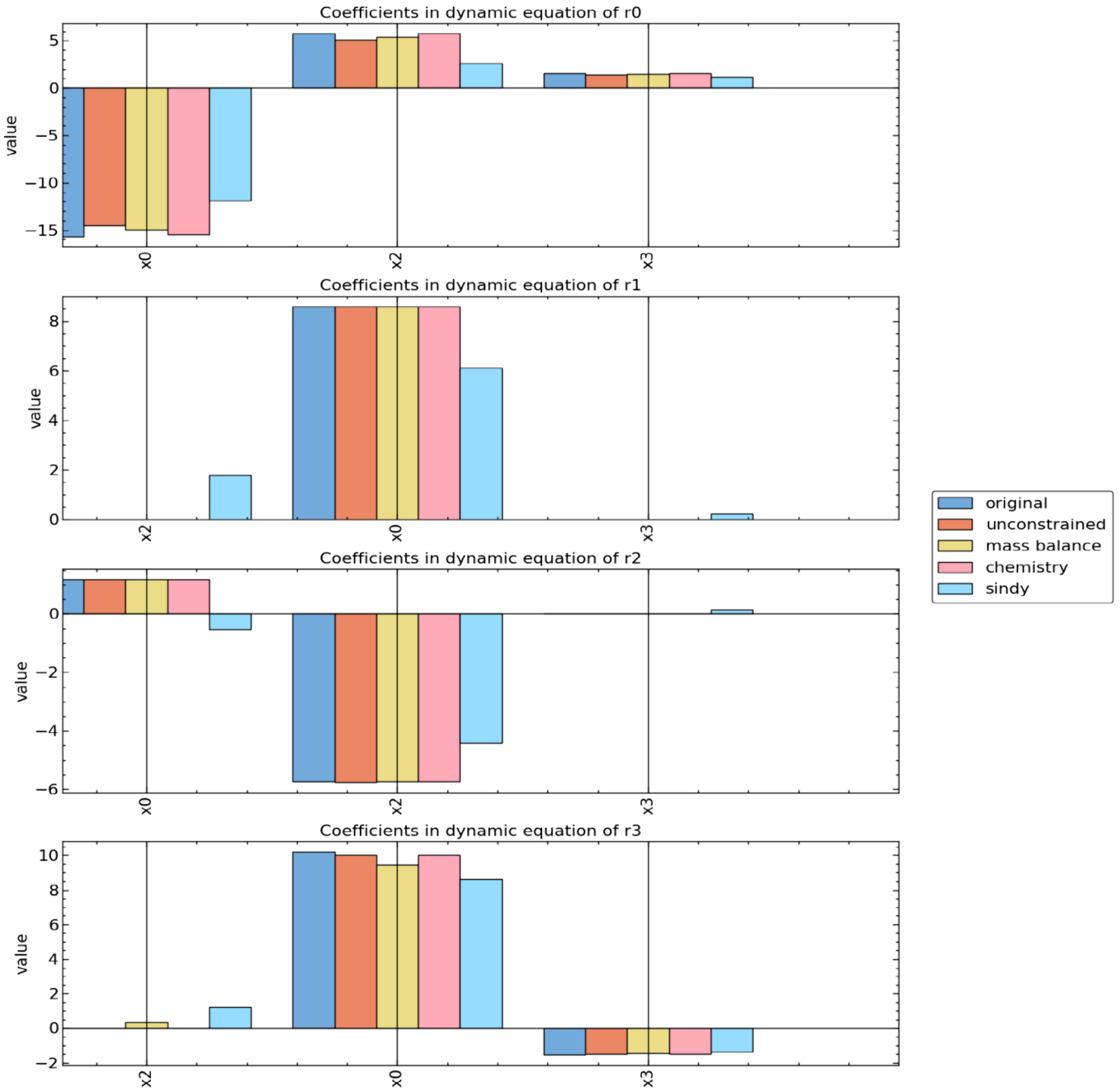


Figure 2. Coefficients and terms of the original model and the discovered models using linear *DF-SINDy* formulations and naive *SINDy* for the case with a Gaussian noise of standard deviation of 0.1. r_{0-3} refer to the rate expression of the consumption/production of species A−D (whose concentrations are termed x_{0-3} in RN1).

Figure 2 and Figure S2 show the coefficient values of nonzero terms in the recovered models in the four formulations vis-a-vis the original (true) model in the presence of noise (in particular, standard deviation of 0.1) for RN1 and RN2 respectively. *SINDy* not only misses a few true terms but also proposes spurious terms. The unconstrained and mass balance formulations of *DF-SINDy* are better than *SINDy* at recovering true terms, but they also propose spurious ones. The chemistry formulation of *DF-SINDy*, in addition to being most accurate in terms of MSE, is also precise in recovering the true terms (i.e., it finds no spurious ones). These results thus reinforce the value of constraining models to as much domain information as possible so that they are quantitatively accurate and qualitatively precise.

Decreasing Number of Experiments. In this subsection, we study the effect of decreasing the amount of data by reducing the number of experiments on the recovery of the model from different formulations. Figure 3a for RN1 and Figure 3b for RN2 compare the performance of unconstrained *DF-SINDy*, the mass balance formulation, and the chemistry-based formulation with naive *SINDy*. We see that the performance of all of the models improves with an increasing number of experiments, as expected. *SINDy* generally performs worse than the three *DF-SINDy* formulations, in terms of either the complexity of the models discovered, MSE, or both. For RN2, *SINDy* fails to recover the original model, while all other formulations accurately identify the correct model when the number of

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

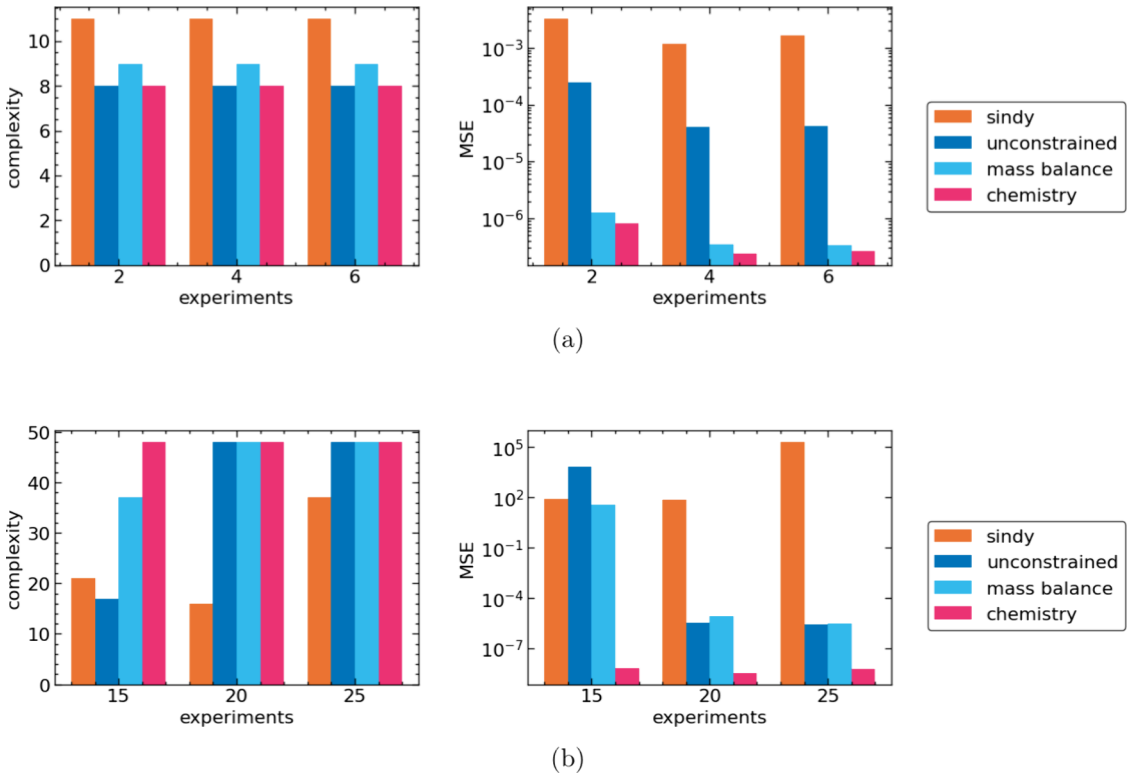


Figure 3. A comparison of model complexity and mean square testing error (MSE) for RN1 (a) and RN2 (b) is shown for the three linear *DF-SINDy* formulations (unconstrained, mass balance, and chemistry) compared with the naive *SINDy* for different number of experiments (i.e., increasing number of runs).

experiments is 20 or higher. This suggests that the integral formulations require fewer data (fewer experiments) to recover accurate models compared to *SINDy*, indicating that integral formulations are more efficient. The chemistry-based formulation consistently performs well for all three levels of data availability and for both reaction networks; as before, it is quantitatively accurate and qualitatively precise. This is also borne out by the recovered terms shown in Figure 4 for RN1; *SINDy* proposes spurious terms and misses the true terms, while the chemistry-based formulation is consistently correct in identifying the true terms (and those alone). This is also observed for RN2 (the coefficient plot in Figure S3).

Decreasing Sampling Frequency. In this subsection, we study the effect of decreasing the amount of data by reducing the sampling frequency on recovering the model with different formulations. We generate training data by varying the output time step $\Delta t = 0.01$, 0.05, and 0.1 (corresponding to a sampling frequency of 100, 20, or 10 samples per second, respectively) while keeping the number of experiments fixed. The performance of the formulations in terms of model complexity and MSE is shown in Figure 5a for RN1 and Figure 5b for RN2; The coefficients and terms of the recovered models for RN1 and RN2 are in Figures S1 and S4. As Δt increases (or frequency decreases), both reaction networks either fail to identify the correct terms or identify coefficients that are slightly off the original coefficients, thereby increasing the MSE of the models. However, again among the formulations, the chemistry-based formulation shows either significantly lower MSE, precise recovery, or both. The unconstrained and mass balance formulations provide intermediate results, while *SINDy* performs the worst.

Nonlinear Formulation: Isothermal Experiments at Different Temperatures. As a final comparison, we consider the problem of learning from different isothermal experiments (conducted at different temperatures) for RN1. Figure 6 shows the performance of the nonlinear formulation with chemistry constraints relative to a modified *SINDy* approach for RN1 and with varying noise. Similar performance for varying sampling frequency and number of experiments is shown in Figure S6 and Figure S8, respectively. In both approaches, we solve a nonlinear optimization problem. In the proposed approach, we solve the differential equations via interpolation and construct the loss on the state measurements while the modified *SINDy* still employs finite differences of the measurements but requires nonlinear optimization to tackle the Arrhenius behavior of rate constants. Both approaches incorporate chemistry information; therefore, the complexity per eq 23 is 10. As can be seen in Figure 6, a chemistry-based nonlinear formulation of *DF-SINDy* consistently performed better than modified *SINDy* by not only identifying the correct terms but also discovering models with low MSE. A comparison of the coefficients of the discovered model with those of the original model is given in Figure S5 for different noise levels, Figure S7 for different numbers of experiments, and Figure S9 for different sampling frequencies.

Effect of Stiffness on Recovery: Michaelis–Menten Kinetics. To show the efficacy of our formulations on stiff systems we consider the Michaelis–Menten reaction network[51] with the reactions given in Table 4. It consists of four different species with one reversible and one irreversible reaction. We consider four different cases, where in each case, the reversible reaction (R_1) is faster than in the previous case. The reaction rates for all cases and each of the species are given in eq 25.

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

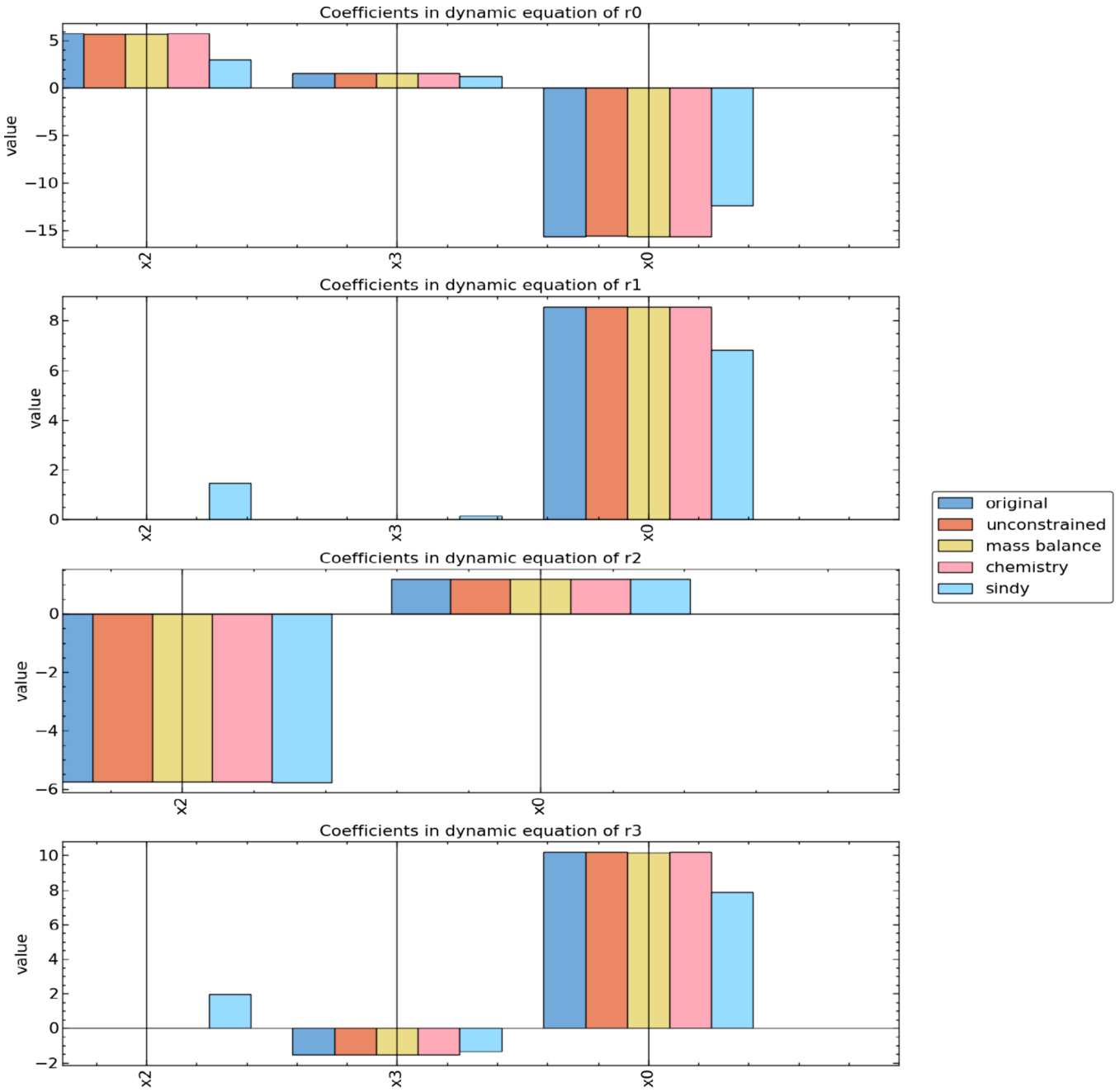


Figure 4. Coefficients and terms of the original model and the discovered models using linear *DF-SINDy* formulations and naive *SINDy* for the case with the number of experiments = 2. r_{0-3} refer to the rate expression of the consumption/production of species A−D (whose concentrations are termed x_{0-3}) in RN1.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ -1 & 1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} k_1AB - k_2C \\ k_3C \end{bmatrix} \tag{25}$$

We observe from Figure 7 that beyond a certain stiffness all formulations fail to identify the true parameters. This is because the reaction dynamics are faster than the sampling frequency ($\Delta t = 0.01$), preventing the capture of accurate concentration transitions. This claim is justified in Figure 8, in which the sampling frequency is increased ($\Delta t = 0.001$) and the recovery is much better, especially for case 4, compared to the models in Figure 7. Furthermore, from Figure S12 and Figure S15, we observe that the integral formulations perform better by identifying parameters that are closer to the true values than the parameters identified using naive *SINDy*. Figures S10, S11, S13, and S14 compare the coefficient plots of the discovered models for the remaining cases. We also observe that even though the model parameters are incorrectly estimated such as in cases 3 and 4 for $\Delta t = 0.01$ and case 4 for $\Delta t = 0.001$, the ratio of forward (k_1) and reverse (k_2) rate constants of the quasi equilibrated species (C) is correctly estimated.

Handling Overinformed Reaction Network. Modern tools such as automated network generators[53] can be used to construct reaction networks comprehensively. However, such tools are more likely to "overestimate" the network, i.e., generate many reactions that allow alternative routes to form the products

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

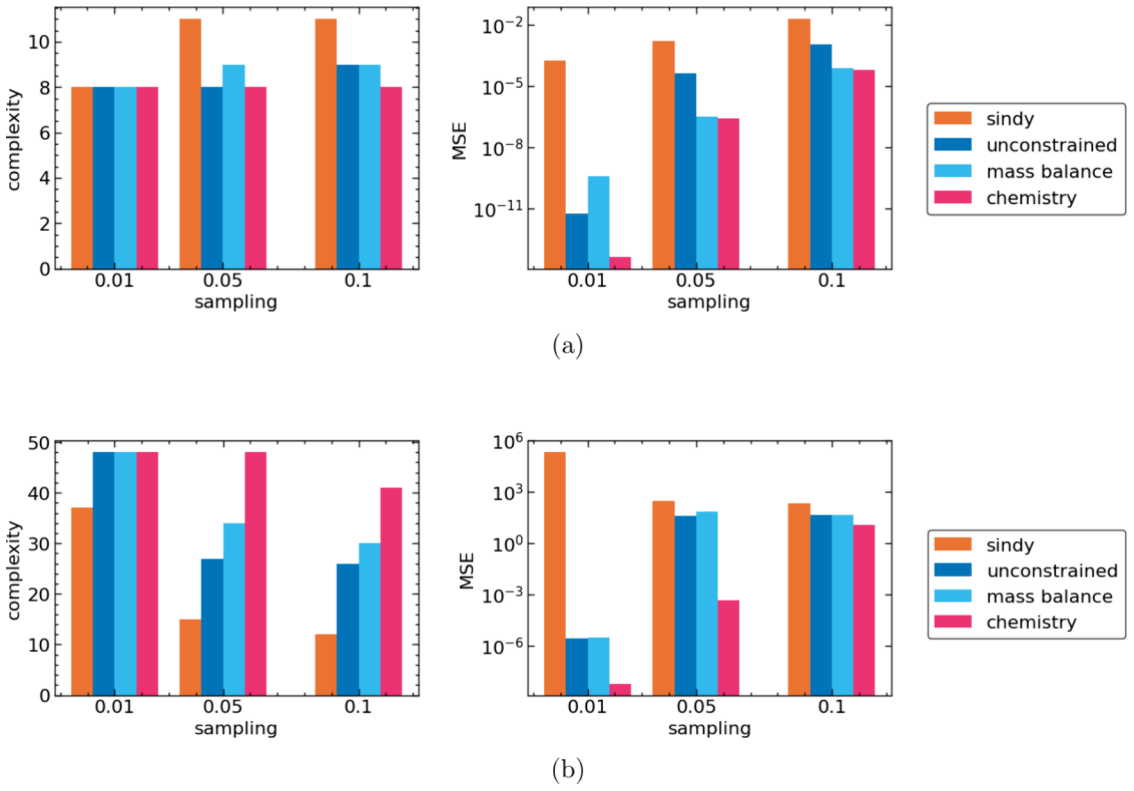


Figure 5. A comparison of model complexity and mean square testing error (MSE) for RN1 (a) and RN2 (b) is shown for the three linear *DF-SINDy* formulations (unconstrained, mass balance, and chemistry) compared with the naive *SINDy* for different sampling frequency (i.e., changing the output time step Δt while generating data).

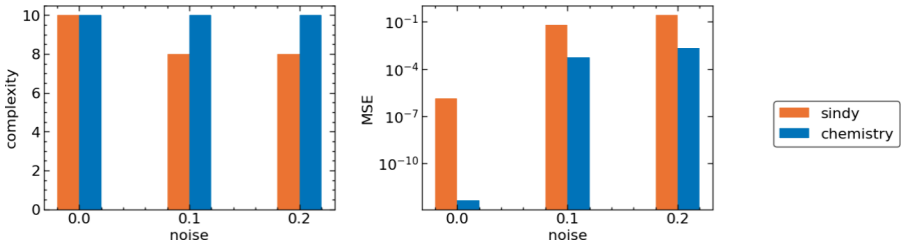


Figure 6. Plot of model complexity and mean squared errors for different noise levels for the nonlinear least-squares problem. The performance of the models discovered from modified *SINDy* and the derivative-free formulation with chemistry constraints for RN2 is compared.

Table 4. Michaelis−Menten Reaction Network with Increasing Stiffness

Case	R_i	Reaction	Forward reaction	Reverse reaction	O(Largest k/Smallest k)
1	1	A + B ↔ C	$k_1 = 0.1$	$k_2 = 0.2$	10^0
	2	C → B + D	$k_3 = 0.3$		
2	1	A + B ↔ C	$k_1 = 1$	$k_2 = 2$	10^1
	2	C → B + D	$k_3 = 0.3$		
3	1	A + B ↔ C	$k_1 = 10$	$k_2 = 20$	10^2
	2	C → B + D	$k_3 = 0.3$		
4	1	A + B ↔ C	$k_1 = 100$	$k_2 = 200$	10^3
	2	C → B + D	$k_3 = 0.3$		

but may ultimately not be flux-carrying. In this section, we show that overestimating the reaction network usually is not a problem, as given enough data, the chemistry formulation of *DF-SINDy* will automatically push the spurious reactions/terms to zero. For the reaction network RN1 given in Table 1, suppose we assume the incorrect reaction mechanism given in Table 5, where the reaction R_1 is reversible and a spurious reversible reaction R_4 is included. We follow the same steps as before but now with the incorrectly specified stoichiometric matrix in eq 26.

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = \begin{bmatrix} -1 & -1 & -1 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \end{bmatrix} \tag{26}$$

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

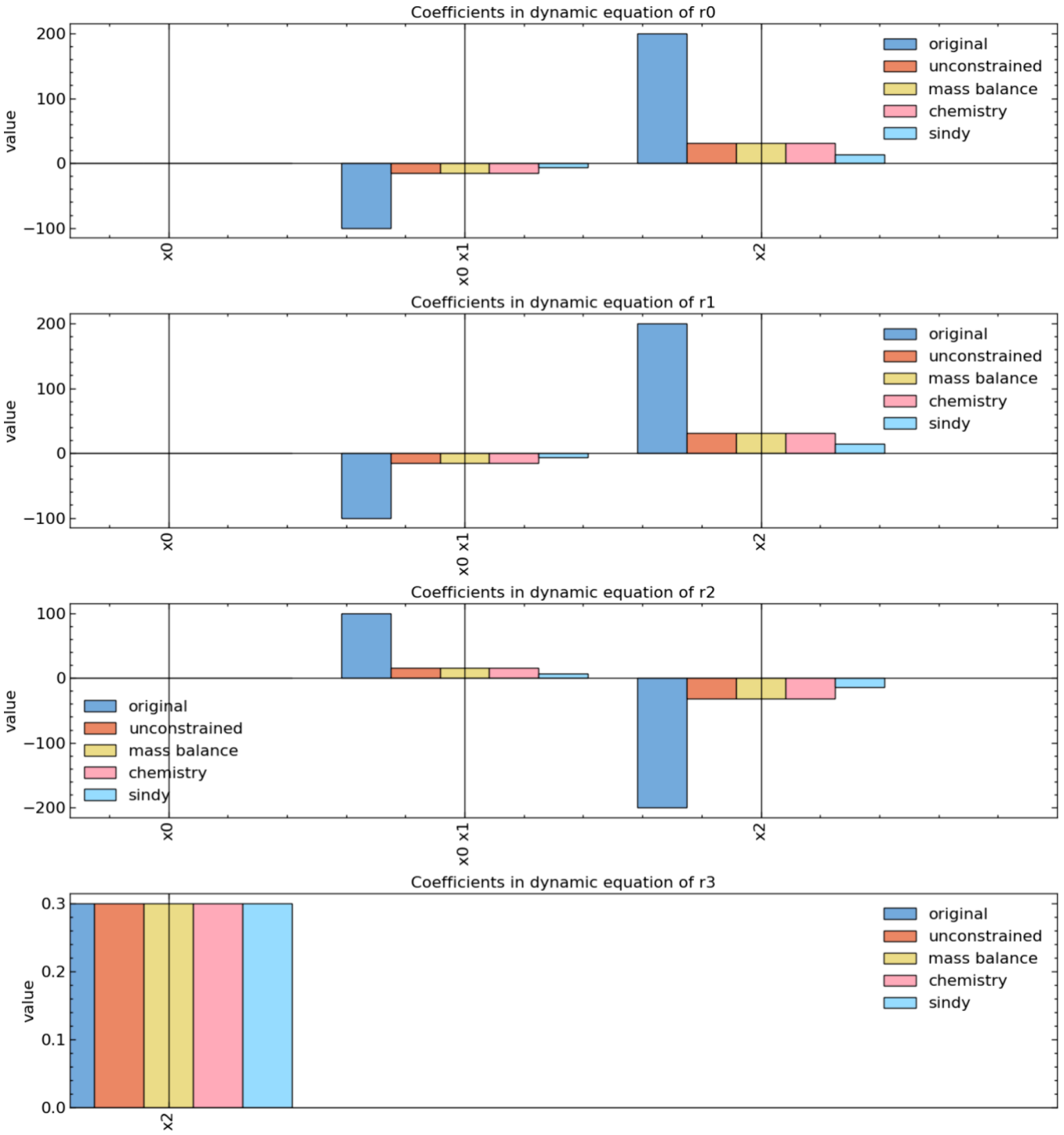


Figure 7. Coefficients and terms of the original model and the discovered models using linear *DF-SINDy* formulations and naive *SINDy* for 4 cases of Table 4 and with sampling frequency $\Delta t = 0.01$. r_{0-3} refer to the rate expression of the consumption/production of species A−D (whose concentrations are termed x_{0-3}).

We compare the coefficients of the discovered model using a linear least-squares formulation (Figure 9) and a nonlinear least-squares formulation (Figure 9) with the coefficients of the original model. We observe that in both cases R_0 is correctly identified as an irreversible reaction (as a single term is discovered) while R_2 and R_3 are reversible reactions (as two terms are discovered). Additionally, spurious reaction R_4 is completely neglected.

CONCLUSION

In this work, we introduced *DF-SINDy*, a method to discover interpretable kinetic models of ordinary differential equations from the reaction data. This method eliminates the need to take derivatives, as was needed in the original *SINDy* method. We further incorporate domain knowledge through mass balance and chemistry formulations, wherein we additionally enforce that the model conserves mass or an underlying reaction network. We show that *DF-SINDy* discovers models that are

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

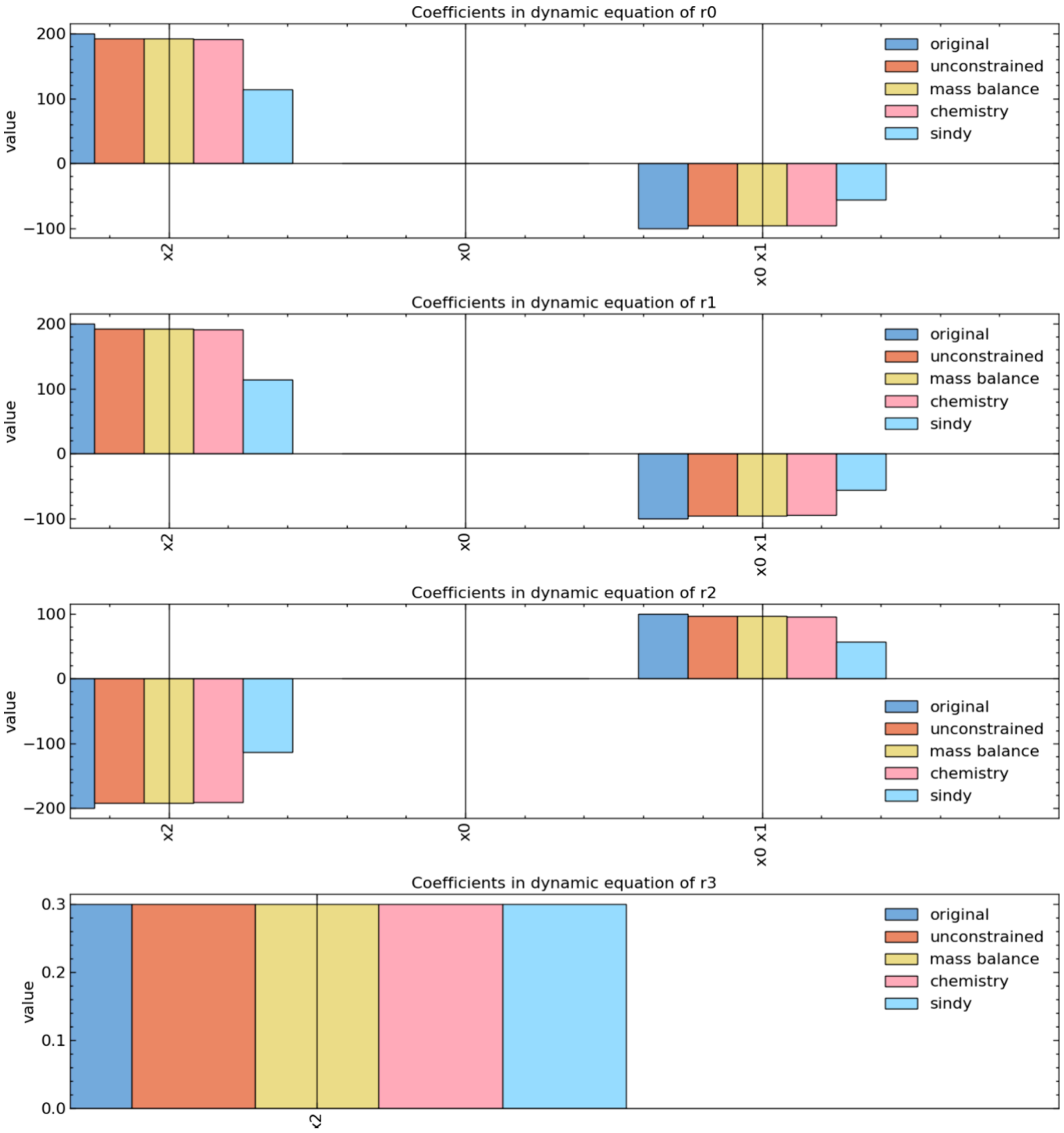


Figure 8. Coefficients and terms of the original model and the discovered models using linear *DF-SINDy* formulations and naive *SINDy* for case 4 of Table 4 and with sampling frequency $\Delta t = 0.001$. r_{0-3} refer to the rate expression of the consumption/production of species A−D (whose concentrations are termed x_{0-3}).

Table 5. Assumed Reaction Mechanism of Cracking and Isomerization of Butenes

R_i	Reaction	Forward reaction	Reverse reaction
1	A ↔ 2B	k_1, E_1	k_2, E_2
2	A ↔ C	k_3, E_3	k_4, E_4
3	A ↔ D	k_5, E_5	k_6, E_6
4	C ↔ D	k_7, E_7	k_8, E_8

quantitatively more accurate (in terms of mean squared errors) and qualitatively more precise (in terms of identifying the correct terms in the model) than those obtained from *SINDy*, especially in the presence of imperfect data, i.e., noisy measurements, less sampling, and fewer experiments. In particular, including chemistry information, i.e., a postulated reaction network, always leads to the best result, wherein the complexity is correctly determined and no spurious terms are identified. Incorporating domain knowledge, thus, (1) reduced

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

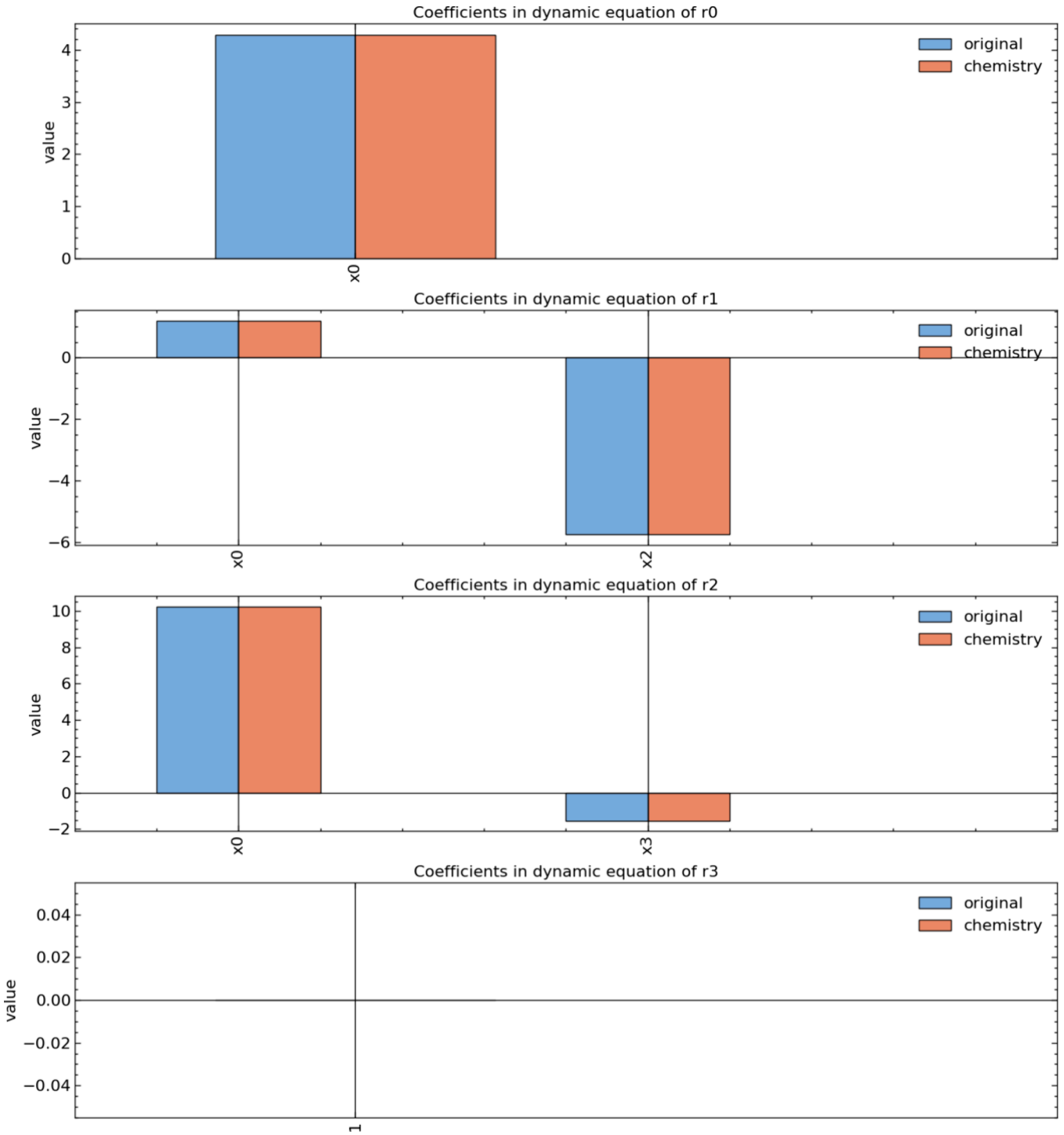


Figure 9. Coefficients and terms of the original model and the discovered models using the derivative-free formulation with overestimated chemistry constraints are compared. r_{0-3} refer to the rate expression of the consumption/production of species A−D (whose concentrations are termed x_{0-3}.

the complexity of the optimization problems, (2) decreased susceptibility to noisy measurements, (3) retained only meaningful terms in the model (especially with chemistry information), and (4) improved interpretability, particularly in inferring underlying reaction mechanisms. Finally, we showed that *DF-SINDy* can be extended to deal with nonlinearity in kinetic parameters, particularly their temperature dependence, although this results in a nonlinear optimization problem. Our results suggest that our approach is quantitatively and qualitatively more accurate than *SINDy*. The biggest advantage of *SINDy*-like methods is the interpretability afforded by the explicit identification of model equations. Therefore, we note that domain-informed *DF-SINDy*, given its relative robustness, is particularly well suited for learning kinetic models from noisy experiments and utilizing them to understand the reaction mechanism (i.e., flux-carrying overall reactions, reaction orders, apparent barriers of steps, etc.) even if elementary steps may not be directly inferred. Since the proposed method works only with measured states, unlike the integration-based nonlinear approaches discussed earlier.[39] While this simplifies the learning

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615

process, it does not directly allow building elementary step models wherein states (e.g., surface intermediates) not measured need to be included, unless domain knowledge or additional measurements (e.g., spectroscopy) allows relating unmeasured states with measured ones via algebraic functions. However, *DF-SINDy* is computationally tractable when the complexity of the reaction system (due to phases, multiplicity of sites, or dynamically evolving catalytic structures) makes it significantly harder to build elementary-step based microkinetic models.

■ ASSOCIATED CONTENT

Data Availability Statement

The code for the algorithms is publicly available on GitHub (https://github.com/siddharth-prabhu/ParameterEstimationConstraints).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.iecr.4c02981.

Additional figures comparing the performance of models obtained from *SINDy* and *DF-SINDy*, coefficient plots for the linear least-squares formulation, coefficient plots for the nonlinear least-squares formulation, and the list of optimal hyperparameters corresponding to the chosen models (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Srinivas Rangarajan − *Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States;* orcid.org/0000-0002-6777-9421; Email: srr516@lehigh.edu

Mayuresh V. Kothare − *Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States*; Email: mvk2@lehigh.edu

Authors

Siddharth Prabhu − *Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States*

Nick Kosir − *Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, Pennsylvania 18015, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.4c02981

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the financial support from Pennsylvania Infrastructure Technology Alliance. S.R. also acknowledges NSF CBET grant 2045550. Portions of this research were conducted on Lehigh University's Research Computing infrastructure partially supported by NSF Award 2019035.

■ REFERENCES

(1) Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 3932−3937.

(2) Motagamwala, A. H.; Dumesic, J. A. Microkinetic Modeling: A Tool for Rational Catalyst Design. *Chem. Rev.* **2021**, *121*, 1049−1076.

(3) Bhandari, S.; Rangarajan, S.; Mavrikakis, M. Combining Computational Modeling with Reaction Kinetics Experiments for Elucidating the In Situ Nature of the Active Site in Catalysis. *Acc. Chem. Res.* **2020**, *53*, 1893−1904.

(4) Matera, S.; Schneider, W. F.; Heyden, A.; Savara, A. Progress in accurate chemical kinetic modeling, simulations, and parameter estimation for heterogeneous catalysis. *ACS Catal.* **2019**, *9*, 6624−6647.

(5) Xie, W.; Xu, J.; Chen, J.; Wang, H.; Hu, P. Achieving theory−experiment parity for activity and selectivity in heterogeneous catalysis using microkinetic modeling. *Acc. Chem. Res.* **2022**, *55*, 1237−1248.

(6) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403−7429.

(7) Wen, M.; Spotte-Smith, E. W. C.; Blau, S. M.; McDermott, M. J.; Krishnapriyan, A. S.; Persson, K. A. Chemical reaction networks and opportunities for machine learning. *Nature Computational Science* **2023**, *3*, 12−24.

(8) Margraf, J. T.; Jung, H.; Scheurer, C.; Reuter, K. Exploring catalytic reaction networks with machine learning. *Nature Catalysis* **2023**, *6*, 112−121.

(9) Mou, T.; Pillai, H. S.; Wang, S.; Wan, M.; Han, X.; Schweitzer, N. M.; Che, F.; Xin, H. Bridging the complexity gap in computational heterogeneous catalysis with machine learning. *Nature Catalysis* **2023**, *6*, 122−136.

(10) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **2017**, *8*, 14621.

(11) Kitchin, J. R. Machine learning in catalysis. *Nature Catalysis* **2018**, *1*, 230−232.

(12) Pablo-García, S.; García-Muelas, R.; Sabadell-Rendón, A.; López, N. Dimensionality reduction of complex reaction networks in heterogeneous catalysis: From l inear-scaling relationships to statistical learning techniques. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e1540.

(13) Chen, Y.-Y.; Ross Kunz, M.; He, X.; Fushimi, R. Recent progress toward catalyst properties, performance, and prediction with data-driven methods. *Curr. Opin. Chem. Eng.* **2022**, *37*, 100843.

(14) Hough, B. R.; Beck, D. A.; Schwartz, D. T.; Pfaendtner, J. Application of machine learning to pyrolysis reaction networks: Reducing model solution time to enable process optimization. *Comput. Chem. Eng.* **2017**, *104*, 56−63.

(15) Ji, W.; Qiu, W.; Shi, Z.; Pan, S.; Deng, S. Stiff-pinn: Physics-informed neural network for stiff chemical kinetics. *J. Phys. Chem. A* **2021**, *125*, 8098−8106.

(16) Gusmao, G. S.; Retnanto, A. P.; Cunha, S. C. d.; Medford, A. J. Kinetics-informed neural networks. *Catal. Today* **2023**, *417*, 113701.

(17) Saadun, A. J.; Pablo-García, S.; Paunovic, V.; Li, Q.; Sabadell-Rendón, A.; Kleemann, K.; Krumeich, F.; López, N.; Pérez-Ramírez, J. Performance of metal-catalyzed hydrodebromination of dibromomethane analyzed by descriptors derived from statistical learning. *ACS Catal.* **2020**, *10*, 6129−6143.

(18) Kunz, M. R.; Yonge, A.; Fang, Z.; Batchu, R.; Medford, A. J.; Constales, D.; Yablonsky, G.; Fushimi, R. Data driven reaction mechanism estimation via transient kinetics and machine learning. *Chemical Engineering Journal* **2021**, *420*, 129610.

(19) Rangarajan, S.; Maravelias, C. T.; Mavrikakis, M. Sequential-Optimization-Based Framework for Robust Modeling and Design of Heterogeneous Catalytic Systems. *J. Phys. Chem. C* **2017**, *121*, 25847−25863.

(20) Gusmao, G. S.; Retnanto, A. P.; Cunha, S. C. d.; Medford, A. J. Kinetics-informed neural networks. *Catal. Today* **2023**, *417*, 113701.

(21) Matera, S.; Schneider, W. F.; Heyden, A.; Savara, A. Progress in Accurate Chemical Kinetic Modeling, Simulations, and Parameter Estimation for Heterogeneous Catalysis. *ACS Catal.* **2019**, *9*, 6624−6647.

(22) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403−7429.
(23) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K.-i. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260−2297.
(24) Wu, Q.; Avanesian, T.; Qu, X.; Van Dam, H. PolyODENet: Deriving mass-action rate equations from incomplete transient kinetics data. *J. Chem. Phys.* **2022**, *157*, 164801.
(25) Hu, P.; Yang, W.; Zhu, Y.; Hong, L. Revealing hidden dynamics from time-series data by ODENet. *J. Comput. Phys.* **2022**, *461*, 111203.
(26) Rangarajan, S.; Tian, H. Improving the predictive power of microkinetic models via machine learning. *Current Opinion in Chemical Engineering* **2022**, *38*, 100858.
(27) Ji, W.; Deng, S. Autonomous Discovery of Unknown Reaction Pathways from Data by Chemical Reaction Neural Network. *J. Phys. Chem. A* **2021**, *125*, 1082−1092.
(28) Schmidt, M.; Lipson, H. Distilling free-form natural laws from experimental data. *science* **2009**, *324*, 81−85.
(29) Hoffmann, M.; Fröhner, C.; Noé, F. Reactive SINDy: Discovering governing reactions from concentration data. *J. Chem. Phys.* **2019**, *150*, 025101.
(30) Mangan, N. M.; Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Inferring Biological Networks by Sparse Identification of Nonlinear Dynamics. *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications* **2016**, *2*, 52−63.
(31) Rudy, S. H.; Brunton, S. L.; Proctor, J. L.; Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **2017**, *3*, No. e1602614.
(32) Champion, K. P.; Brunton, S. L.; Kutz, J. N. Discovery of Nonlinear Multiscale Systems: Sampling Strategies and Embeddings. *SIAM Journal on Applied Dynamical Systems* **2019**, *18*, 312−333.
(33) Jiang, R.; Singh, P.; Wrede, F.; Hellander, A.; Petzold, L. Identification of dynamic mass-action biochemical reaction networks using sparse Bayesian methods. *PLoS Comput. Biol.* **2022**, *18*, e1009830.
(34) Bhatt, N.; Jayawardhana, B.; Plaza, S. SINDy-CRN: Sparse Identification of Chemical Reaction Networks from Data. *Proc. IEEE Conf. Decis. Control* **2023**, 3512−3518, DOI: 10.1109/CDC49753.2023.10384032.
(35) Wentz, J.; Doostan, A. Derivative-based SINDy (DSINDy): Addressing the challenge of discovering governing equations from noisy data. *Computer Methods in Applied Mechanics and Engineering* **2023**, *413*, 116096.
(36) Van Breugel, F.; Kutz, J. N.; Brunton, B. W. Numerical Differentiation of Noisy Data: A Unifying Multi-Objective Optimization Framework. *IEEE Access* **2020**, *8*, 196865−196877.
(37) Chartrand, R. Numerical differentiation of noisy, nonsmooth data. *ISRN Applied Mathematics* **2011**, *2011*, 1−11.
(38) Kaheman, K.; Brunton, S. L.; Kutz, J. N. Automatic differentiation to simultaneously identify nonlinear dynamics and extract noise probability distributions from data. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 015031.
(39) Lejarza, F.; Koninckx, E.; Broadbelt, L. J.; Baldea, M. A dynamic nonlinear optimization framework for learning data-driven reduced-order microkinetic models. *Chemical Engineering Journal* **2023**, *462*, 142089.
(40) Lejarza, F.; Baldea, M. Data-driven discovery of the governing equations of dynamical systems via moving horizon optimization. *Sci. Rep.* **2022**, *12*, 11836.
(41) Hirsh, S. M.; Barajas-Solano, D. A.; Kutz, J. N. Sparsifying priors for Bayesian uncertainty quantification in model discovery. *Royal Society Open Science* **2022**, *9*, 211823.
(42) Fasel, U.; Kutz, J. N.; Brunton, B. W.; Brunton, S. L. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proc. R. Soc. A* **2022**, *478*, 20210904.
(43) Joshi, M.; Seidel-Morgenstern, A.; Kremling, A. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering* **2006**, *8*, 447−455.
(44) Schaeffer, H.; McCalla, S. G. Sparse model selection via integral terms. *Phys. Rev. E* **2017**, *96*, 023302.
(45) Naber, M.; Haeseler, F. v.; Rudolph, N.; Huber, H. J.; Findeisen, R. Parameter Estimation by Picard-Iteration for Biochemical Networks with Noisy Data. *IFAC-PapersOnLine* **2018**, *51*, 64−67.
(46) Calver, J. Parameter Estimation for Systems of Ordinary Differential Equations. Ph.D. thesis, University of Toronto, 2019.
(47) de Boor, C. *A Practical Guide to Splines*; Springer: New York, NY, 2001.
(48) Wächter, A.; Biegler, L.; Lang, Y.-d.; Raghunathan, A. *IPOPT: An interior point algorithm for large-scale nonlinear optimization*, 2002. https://github.com/coin-or/Ipopt.
(49) Andersson, J. A. E.; Gillis, J.; Horn, G.; Rawlings, J. B.; Diehl, M. CasADi − A software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation* **2019**, *11*, 1−36.
(50) Gupta, U.; Heo, S.; Bhan, A.; Daoutidis, P. Time scale decomposition in complex reaction systems: A graph theoretic analysis. *Comput. Chem. Eng.* **2016**, *95*, 170−181.
(51) Vora, N.; Daoutidis, P. Nonlinear model reduction of chemical reaction systems. *AIChE J.* **2001**, *47*, 2320−2332.
(52) Fajardo-Fontiveros, O.; Reichardt, I.; De Los Ríos, H. R.; Duch, J.; Sales-Pardo, M.; Guimerà, R. Fundamental limits to learning closed-form mathematical models from data. *Nat. Commun.* **2023**, *14*, 1043.
(53) Rangarajan, S.; Bhan, A.; Daoutidis, P. Language-oriented rule-based reaction network generation and analysis: Description of RING. *Comput. Chem. Eng.* **2012**, *45*, 114−123.

https://doi.org/10.1021/acs.iecr.4c02981
Ind. Eng. Chem. Res. 2025, 64, 2601−2615