

# Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: M. A. Palma Banos, J. D. Kress, R. Hernandez and G. T. Craven, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00470E.



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

Cite this: DOI: 00.0000/xxxxxxxxxx

# Stoichiometrically-informed symbolic regression for extracting chemical reaction mechanisms from data

Manuel Palma Banos,<sup>a</sup> Joel D. Kress,<sup>b</sup> Rigoberto Hernandez,<sup>a</sup> and Galen T. Craven,<sup>b</sup>Received Date  
Accepted Date

DOI: 00.0000/xxxxxxxxxx

A data-driven computational method is introduced to extract chemical reaction mechanisms from time series chemical concentration data. It is realized through the use of dynamic symbolic regression in which a sparse analytical form for a dynamical system is discoverable from the underlying data. We specifically develop the stoichiometrically-informed symbolic regression (SISR) method to address a standing challenge in complex chemical reaction networks: Given a time-series dataset of concentrations of several components, what is the mechanism and the associated rate constants? SISR finds the optimal mechanism, kinetic equations and rate constants by combining differential optimization with a genetic optimization approach that searches a symbolic space of possible reaction mechanisms. Use of SISR in several paradigmatic examples spanning linear and nonlinear reaction schemes results in excellent agreement between true and predicted mechanisms, including when the method is applied to noisy data. The advantages of a stoichiometrically-informed approach such as SISR to address reaction discovery is illustrated through comparison with the use of generic state-of-the-art data-driven approaches.

## 1 Introduction

Determining chemical reaction mechanisms is foundational in many research areas such as catalysis, electrochemistry, combustion, and biochemistry that feature prominently in modern scientific and technological landscapes<sup>1–10</sup>. Chemical reaction mechanisms give fundamental insight into a physicochemical process, providing elucidation and allowing interpretation of the underlying chemical reactions that give rise to a process. Chemical mechanisms can also be used to forecast how the outcome or output of a process will change over time. However, deriving a set of kinetic mechanistic equations that accurately describes the time evolution of concentrations of chemical species involved in a mechanism is often difficult or simply intractable in practice due to, for example, complex nonlinear interactions between reacting species, a large number of chemical species participating in the process, and/or reactions occurring over multiple timescales. Deriving chemical reaction mechanisms by hand generally requires physical intuition about a system and subject matter expertise<sup>11,12</sup>. This is because determining accurate functional forms for reaction mechanisms typically involves searching a vast space of possible reactions that are involved in a process while also determining how those reactions are coupled in the overall mechanism. This problem is compounded because not only does the

reaction mechanism itself need to be determined, but the chemical rate constants describing species-to-species transformations, i.e., chemical reactions, must be parameterized, often over varying thermodynamic conditions such as different temperatures and pressures<sup>13–17</sup>.

Because of these difficulties, automated reaction mechanism generation is an emerging data-driven research approach that can accelerate the extraction of accurate reaction mechanisms from data<sup>5–8,18–20</sup>. Data-driven and machine learning (ML) methods have been broadly and successfully applied in many areas of the physical sciences<sup>21–35</sup>. Data-driven approaches for reaction discovery have been used to decipher complex and large datasets of chemical concentration data by extracting chemical reaction pathways, rate constants, and reaction mechanisms<sup>5–8,36</sup>. The current data-driven reaction mechanism discovery methods, however, generally suffer from the same shortcomings that are typical of most data-driven and ML approaches including lacking interpretability, a large number of parameters, the black-box nature of the approximating function, and poor performance when extrapolation outside of the training data is performed. One ML approach that is used to circumvent these limitations is Symbolic Regression (SR)—a method to search for simple analytical functions that best describe a dataset. SR has been applied in multiple contexts to extract sparse and interpretable functional forms from data<sup>37–39</sup>. In the context of SR applied to dynamical systems<sup>40</sup>, dynamical SR approaches can be used to extract a sparse analytical form for a dynamical system from time series data. Integrat-

<sup>b</sup> Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.<sup>a</sup> Department of Chemistry, Johns Hopkins University, Baltimore, Maryland, USA.

ing the discovered system of dynamical equations will generate the time-evolution of the input variables, for example the time-dependence of the concentrations of chemical species<sup>41,42</sup>.

One of the most prominent SR methods that is used to discover the dynamical equations giving rise to a time series dataset is the Sparse Identification of Nonlinear Dynamical systems (SINDy) approach<sup>40,43,44</sup>. SINDy has been applied broadly in the physical sciences and has seen successful applications in diverse areas such as biological networks<sup>45–47</sup>, aerodynamics<sup>48</sup>, and plasma physics<sup>49</sup>, among others. One of the primary advantages of SINDy is that given a collection of time series data, it can quickly (relative to, for example, genetic SR approaches<sup>50,51</sup>) search a large space of possible analytical functional forms and corresponding parameter values to generate a simple dynamical system that when integrated matches the time evolution of the input data. Other approaches to determine symbolic expressions for physical mechanisms have been developed. For example, the sure-independence screening and sparsifying operator (SISSO) method<sup>52,53</sup> can be applied to generate sparse analytical descriptors of a material's properties.

The SINDy framework has been applied to determine chemical reaction mechanisms using an approach termed Reactive SINDy (see Ref. 41). This interesting application of SR to chemical reaction networks produced sparse reaction mechanisms in good agreement with the input data and illustrates the potential of dynamical SR in macroscale chemical reaction dynamics. The Reactive SINDy approach does have limitations that reduce its utility and robustness including: (1) The user must propose a collection of reaction ansatz, i.e., the user must guess what reactions are present in a process. This requirement can be cumbersome and time-consuming especially if the number of species being studied is large or little is known about the physical process being examined. (2) There are no constraints on the reaction rates in the derived mechanism which can lead to unphysical results such as negative concentrations when the system is integrated. (3) Fast-slow dynamics<sup>54</sup> are not well-described. For example, if there are rate constants that differ by multiple orders of magnitude the SINDy approach will generally prune the slow process from the derived dynamical system. This poses a problem because elimination of reactions with small rate constants can eliminate reaction pathways that are vital to the overall mechanism. While these problems are not present in all chemical reaction networks/mechanisms, they do limit the applicability and utility of Reactive SINDy in some cases<sup>41</sup>. Other approaches such as SINDy - CRN<sup>42</sup> where CRN stands for Chemical Reaction Network, and the one defined in Ref. 55 seek to alleviate some of these problems.

Other methodologies for extracting dynamical systems from data have been developed and applied to good effect<sup>56–63</sup>. In the context of chemical reaction mechanism discovery, it would be advantageous to developing a method that (a) gives the explicit individual reactions involved in a process, (b) gives the stoichiometry of those reactions—a fundamental property in the analysis of chemistry and chemical reactions, (c) does not rely on neural network formulations of the chemical reaction network, as they can reduce interpretability and accurate extrapolation (i.e., ac-

curate time-series forecasting), (d) can detect hidden variables such as unknown chemical intermediates in a reaction mechanism, (e) does not require a postulated set of potential reactions to be included as reaction ansatz, (f) accurately returns the rate constants for each reaction, and (g) is robust to noise in data.

In this work, we develop and apply a stoichiometrically informed symbolic regression (SISR—pronounced “scissor”) tool to determine chemical reaction mechanisms and chemical kinetic equations from time-series concentration data. Our specific technical advance is to apply a physics-informed mathematical formalism that accounts for intrinsic stoichiometry in a chemical reaction to automate the discovery of accurate chemical reaction mechanisms from data. The developed method returns sparse and interpretable analytical forms for a reaction mechanism discovered from data. A genetic optimization approach is employed to search the symbolic space of possible reaction mechanisms to find the one that best matches a time-series dataset of chemical concentrations. That genetic approach is coupled with the stoichiometrically-informed method to fit the rate constants in a reaction mechanism through differential optimization. Applying the method results in excellent agreement between true and predicted mechanisms over data from multiple linear and nonlinear reaction schemes. The agreement is shown to persist over sparse and noisy datasets, such as those that would typically be obtained from experiments.

The remainder of this article is organized as follows: Section 2 contains details of methods and the formalism that are applied including the genetic search procedure over the symbolic reaction space and the numerical procedures used to fit the rate constants in those reactions. In Sec. 3, the results of the method for several model reactive schemes and chemical reaction networks are presented. Conclusions and future directions are discussed in Sec. 4.

## 2 Stoichiometrically-Informed Symbolic Regression (SISR)

### 2.1 Data Structure and Mathematical Formalism

The overall goal of SISR is to take time series concentration data for chemical processes where the underlying reactions are unknown, and to extract the correct reactions and rate constants from that data. The SISR method is described using the following mathematical formalism. Consider a dataset of chemical concentration data containing  $N$  chemical species, where the concentration of each species is measured at times  $t_1, t_2, \dots, t_m$ . Expressed in matrix form where each column is the times series concentration data for a different species, this dataset is

$$S = \begin{bmatrix} [S_1](t_1) & [S_2](t_1) & \cdots & [S_N](t_1) \\ [S_1](t_2) & [S_2](t_2) & \cdots & [S_N](t_2) \\ \vdots & \vdots & \ddots & \vdots \\ [S_1](t_m) & [S_2](t_m) & \cdots & [S_N](t_m) \end{bmatrix}$$

where the  $[S_k]$  notation represents the concentration of chemical species  $S_k$ . The goal is to derive a *symbolic* reaction mechanism  $\mathbf{M}$  and the corresponding set of rate constants  $\mathbf{k}$  that best fits the dataset  $S$ . The total reaction mechanism comprises a set of



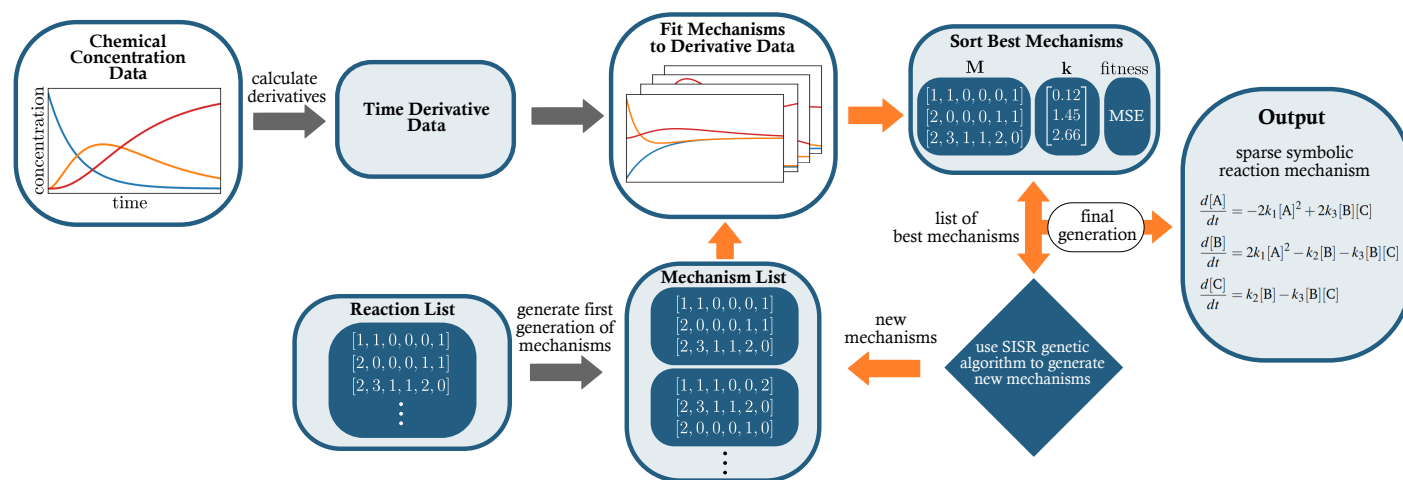


Fig. 1 Schematic diagram showing the workflow for the developed SISR method.

chemical reactions involved in the process, the rate constants for those reactions, and the corresponding set of kinetic equations for the set of reactions. Our approach is to use a genetic algorithm to evolve stoichiometry-constrained symbolic expressions for a collection of reaction mechanisms, fit the rate constants in those mechanisms to numerical derivatives of the concentration data, and repeat this process for multiple iterations (generations) using the best mechanisms from the previous generation to generate the next generation. After a set number of generations are evolved, a final determination of the best overall mechanism is made. A schematic diagram of the developed workflow is shown in Fig. 1.

The mechanisms are fit and constructed in the derivative space of the concentration data. Here, the numerical derivatives of the concentration data, obtained using finite difference methods, are represented by

$$\dot{S} = \begin{bmatrix} \left. \frac{d[S_1]}{dt} \right|_{t=t_1} & \left. \frac{d[S_2]}{dt} \right|_{t=t_1} & \dots & \left. \frac{d[S_N]}{dt} \right|_{t=t_1} \\ \left. \frac{d[S_1]}{dt} \right|_{t=t_2} & \left. \frac{d[S_2]}{dt} \right|_{t=t_2} & \dots & \left. \frac{d[S_N]}{dt} \right|_{t=t_2} \\ \vdots & \vdots & \ddots & \vdots \\ \left. \frac{d[S_1]}{dt} \right|_{t=t_m} & \left. \frac{d[S_2]}{dt} \right|_{t=t_m} & \dots & \left. \frac{d[S_N]}{dt} \right|_{t=t_m} \end{bmatrix}$$

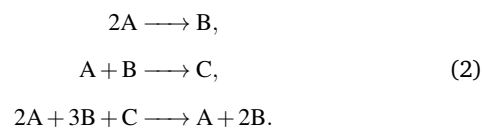
Fitting to derivatives is a common approach used to discover symbolic dynamical systems from data because it allows the derivative functions, for example,  $\frac{dx}{dt}$ ,  $\frac{dy}{dt}$ ,  $\frac{dz}{dt}$ , to be constructed directly as opposed to constructing the primary functions,  $x(t)$ ,  $y(t)$ ,  $z(t)$ , and then deriving the dynamical system from those functions<sup>40,64</sup>. We apply this same approach.

A symbolic reaction mechanism **M** is a collection of chemical reactions

$$\mathbf{M} = [\text{rxn}_i, \text{rxn}_j, \text{rxn}_k, \dots], \quad (1)$$

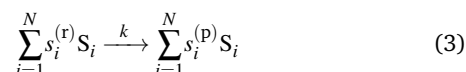
where each reaction (rxn) is associated with a rate constant  $k_{\text{rxn}}$ . The total mechanism is a combination of the symbolic mechanism **M** and an array of rate constants **k** that contains a numerical value for the rate constant of each reaction in that mechanism. For example, some reactions that are possible in a process involving

three chemical species  $S_1 = A$ ,  $S_2 = B$ , and  $S_3 = C$  are



The size of a mechanism is given by its cardinality  $|\mathbf{M}|$  which is the number of reactions in the mechanism.

To illustrate the SISR method, consider the abstract chemical reaction example



where the terms on the LHS are reactants (denoted by the superscript “r”) and the terms on the RHS are products (denoted by the superscript “p”). The sums are taken over all the chemical species involved in the process. The stoichiometric coefficients of species  $S_i$  in the reactant and product states are  $s_i^{(r)}$  and  $s_i^{(p)}$ , respectively. To represent each reaction in vector form (a convenient mathematical notation for our purposes), we break each reaction into a reactant vector and a product vectors, each containing the stoichiometric coefficients in the respective state. The reactant vector for Eq. (3) is

$$\text{rxn}^{(r)} = [s_1^{(r)}, s_2^{(r)}, \dots, s_N^{(r)}], \quad (4)$$

the product vector is

$$\text{rxn}^{(p)} = [s_1^{(p)}, s_2^{(p)}, \dots, s_N^{(p)}], \quad (5)$$

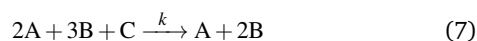
and the total reaction vector is obtained through a concatenation of the reactant and product vectors:

$$\begin{aligned} \text{rxn} &= [s_1^{(r)}, s_2^{(r)}, \dots, s_N^{(r)}] \oplus [s_1^{(p)}, s_2^{(p)}, \dots, s_N^{(p)}] \\ &= [s_1^{(r)}, s_2^{(r)}, \dots, s_N^{(r)}, s_1^{(p)}, s_2^{(p)}, \dots, s_N^{(p)}] \in \mathbb{N}_0^{2N}, \end{aligned} \quad (6)$$

where  $\oplus$  represents the concatenation operation. For example, consider again a reaction involving three species:  $S_1 = A$ ,  $S_2 =$



B, and  $S_3 = C$  defined by



The reactant vector for this reaction is

$$\text{rxn}^{(r)} = [2, 3, 1], \quad (8)$$

the product vector is

$$\text{rxn}^{(p)} = [1, 2, 0], \quad (9)$$

and the total reaction vector is

$$\text{rxn} = [2, 3, 1] \oplus [1, 2, 0] = [2, 3, 1, 1, 2, 0]. \quad (10)$$

Using this formalism, chemical reactions can be expressed in convenient mathematical form that maps to a symbolic representation for analysis and optimization.

Each total mechanism has a corresponding system of symbolic kinetic equations defined by:

$$\begin{aligned} \frac{d[S_1]}{dt} &= \sum_{\text{rxn} \in \mathbf{M}} k_{\text{rxn}} \left( s_1^{(p)} - s_1^{(r)} \right) \prod_{j=1}^N [S_j]^{s_j^{(r)}}, \\ \frac{d[S_2]}{dt} &= \sum_{\text{rxn} \in \mathbf{M}} k_{\text{rxn}} \left( s_2^{(p)} - s_2^{(r)} \right) \prod_{j=1}^N [S_j]^{s_j^{(r)}}, \\ &\dots \\ \frac{d[S_N]}{dt} &= \sum_{\text{rxn} \in \mathbf{M}} k_{\text{rxn}} \left( s_N^{(p)} - s_N^{(r)} \right) \prod_{j=1}^N [S_j]^{s_j^{(r)}}. \end{aligned} \quad (11)$$

The kinetic equations are integrated to determine the time evolution of the concentration of each species.

## 2.2 Genetic Search for the Optimal Reaction Mechanism

We now describe the mathematics of how symbolic mechanisms are evolved in SISR.

### 2.2.1 Reaction List

The first step is to create a reaction list—a list of all possible reactions that could be included in a mechanism. The reaction list is created based on predefined constraints for the *maximum reaction order*,  $O$ , and *maximum stoichiometric ratio*,  $R$ , that can be included in each reaction on the list. Each reaction in a mechanism is represented by a reaction vector like in Eq. (6) with  $2N$  elements, where the first  $N$  elements are the stoichiometric coefficients of the reactants, the sum of which must satisfy the constraint

$$\sum_{i=1}^N s_i^{(r)} \leq O. \quad (12)$$

This is a mathematical statement that a reaction on the reaction list must not exceed the predefined reaction order. Typical values for  $O$  will be  $O = 1$  if only first order (and optionally zeroth order) reactions are to be included or  $O = 2$  if first and second order reactions (and optionally zeroth order) are to be included, although higher reaction orders can also be used. The last  $N$  elements in the reaction vector are the stoichiometric coefficients of

the products, the sum of which in the SISR method is constrained by:

$$\sum_{i=1}^N s_i^{(p)} \leq O * R. \quad (13)$$

The number of possible reactant vectors is given by  $N + O$ . Choose  $O$ , i.e.,  $\binom{N+O}{O}$ , and the number of possible product vectors is  $(O * R)^N$ . The final list of reactions is formed by combining every possible reactant vector with all possible product vectors and eliminating redundant or stoichiometrically prohibited reactions.

We utilize an *islanding* procedure in the SISR genetic algorithm, where multiple islands are created, each containing mechanisms with a fixed number of reactions,  $|\mathbf{M}|$ . Islanding is a technique used in genetic algorithms where the population is divided into distinct subpopulations (here based on the number of reactions in a mechanism), each evolving separately on different "islands". The mechanisms in each island maintain a constant number of reactions throughout the evolutionary process. The genetic algorithm is then applied separately to each island, allowing independent evolution of mechanisms with their respective constraints on the number of reactions. No information is transferred between islands during the search procedure. The islanding approach promotes diversity among potential solutions and mitigates the need to address mechanism complexity (or other factors) for optimal mechanism selection during the search on each island. This ultimately leads to more efficient optimization and also makes the search easier to parallelize.

### 2.2.2 First generation

In the next step, the first generation of possible reaction mechanisms

$$\mathbf{G}_1 = [\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \dots], \quad (14)$$

is constructed for each island using the reaction list. In the first generation,  $n_{\text{mech}}$  mechanisms are constructed. Each mechanism is constructed by randomly and sequentially selecting reactions from the reaction list until (a) all the chemical species that are involved in a process are included in that mechanism (either as reactant, product, or both) and (b) the number of reactions in the mechanism is equal to the cardinality of the specific island size. The first constraint is imposed so that if the concentration data set contains data for  $N$  chemical species, then every mechanism should involve  $N$  chemical species, i.e., if  $N$  chemical species are represented in the data, then all  $N$  species must be involved in the mechanism. Each reaction in the reaction list is equally weighted in this initial selection procedure. The outcome is collection of  $n_{\text{mech}}$  possible mechanisms.

After constructing symbolic expressions for each reaction mechanisms in the generation, the next step is to fit the rate constants for each mechanism to the data. Our aim is to minimize the discrepancy between the true concentration derivatives  $\frac{d[S_i]}{dt}$  and the predicted derivatives  $\frac{d[\hat{S}_i]}{dt}$  for each species  $i$  where the hat notation signifies that the derivatives arise from fits to the data. The time derivatives are computed using a second-order accurate central difference for interior points and first-order accurate forward/backward differences at the boundaries. This approach is able to handle both uniformly-spaced and nonuniformly-spaced



time-series datasets.

The rate constant fitting is achieved using the mean squared error (MSE) defined through the loss function in the derivative space

$$\mathcal{L}_{\text{der}} \equiv \frac{1}{Nm} \sum_{j=1}^m \sum_{i=1}^N \left( \frac{\left. \frac{d[S_i]}{dt} \right|_{t=t_j}}{\max \left( \left| \frac{d[S_i]}{dt} \right| \right)} - \frac{\left. \frac{d[\hat{S}_i]}{dt} \right|_{t=t_j}}{\max \left( \left| \frac{d[\hat{S}_i]}{dt} \right| \right)} \right)^2, \quad (15)$$

as an overall loss metric for each mechanism where the sum over  $j$  accounts for all the time points in the dataset and the sum over  $i$  accounts for all the chemical species. Each derivative value in the MSE is scaled so that species with large derivative value do not dominate the error calculation. The fitting process is performed by finding the set of rate constants that minimize the MSE between the observed and predicted derivatives:

$$\mathbf{k}_{\text{fit}} = \underset{\mathbf{k}}{\text{argmin}}(\mathcal{L}_{\text{der}}(\mathbf{k})). \quad (16)$$

The time derivatives for each species in a mechanism are fit to expressions that encode stoichiometric information in the form:

$$\left. \frac{d[\hat{S}_i]}{dt} \right|_{t=t_j} = \sum_{\text{rxn} \in \mathbf{M}} k_{\text{rxn}} (s_i^{(p)} - s_i^{(r)}) \prod_{j=1}^N [S_j](t_j)^{s_j^{(r)}}, \quad (17)$$

for each species. The minimization is performed using nonlinear least squares regression implemented through the trust region reflective method. All of the rate constants are fit at the same time, as opposed to a sequential fitting procedure. After numerical values for the rate constants in every mechanism have been assigned using the fitting procedure, we have the first generation of fit mechanisms

$$\mathbf{G}_1^{(\text{fit})} = [\mathbf{M}_1^{(\text{fit})}, \mathbf{M}_2^{(\text{fit})}, \mathbf{M}_3^{(\text{fit})}, \dots], \quad (18)$$

and a corresponding set of values for the loss function

$$\mathcal{L}_1 = [\mathcal{L}_{\text{der}}^{(1)}, \mathcal{L}_{\text{der}}^{(2)}, \mathcal{L}_{\text{der}}^{(3)}, \dots]. \quad (19)$$

The initial generation of mechanisms is then sorted and ranked based on fitness defined by the MSE in the derivative space, i.e.,  $\mathcal{L}_{\text{der}}$ .

### 2.2.3 Next generations

The  $n_{\text{best}} = \text{int}(\mathcal{E}n_{\text{mech}})$  fittest mechanisms from the previous generation  $\mathbf{G}_i$  are kept for the next generation  $\mathbf{G}_{i+1}$  using an elitism fraction  $\mathcal{E}$ . Therefore, in each subsequent generation after the first,  $n_{\text{new}} = n_{\text{mech}} - n_{\text{best}}$  new mechanisms must be created using information from the best mechanisms from the previous generation. This process is performed using crossover methods and then mutation methods.

Crossover involves taking information (reactions, reactant vectors, and/or product vectors) from the best performing mechanisms and using that information to generate new mechanisms. The probability of a mechanism from the previous generation being involved in a crossover event comes from the ranked-based

selection:

$$p_i = \frac{n_{\text{mech}} - i + \varepsilon}{\sum_{j=1}^{n_{\text{mech}}} (n_{\text{mech}} - j + \varepsilon)}, \quad (20)$$

where  $\varepsilon$  is a small fractional numerical value (throughout taken to be 0.2) used so that a non-zero probability is assigned to all the mechanisms in a generation. We use this ranking procedure instead of weighting directly according to MSE value to avoid high-fitness mechanisms dominating in early generations. The ranked-based weighting adds diversity to the pool of solutions.

The crossover mechanism is generated by choosing two mechanisms (the parents) randomly according to the probability  $p_i$ . The reactions contained in the two selected parent mechanisms are then combined into a *gene pool* which is a collection of all the reactions involved in the chosen mechanisms. The gene pool is edited so reactions only appear once. Two offspring mechanisms are then created by randomly selecting reactions from this *gene pool*. The offspring are created by randomly and sequentially selecting reactions from the gene pool until all the chemical species that are involved in a process are included in children mechanisms and the number of reactions in the mechanism is equal to the cardinality of the specific island size being evolved. For example, consider the two parent mechanisms from an  $|\mathbf{M}| = 4$  island:

$$\mathbf{M}_1 = [\text{rxn}_1, \text{rxn}_2, \text{rxn}_3, \text{rxn}_4], \quad (21)$$

$$\mathbf{M}_2 = [\text{rxn}_5, \text{rxn}_6, \text{rxn}_7, \text{rxn}_8]. \quad (22)$$

The gene pool for these parents is

$$\text{GP} = [\text{rxn}_1, \text{rxn}_2, \text{rxn}_3, \text{rxn}_4, \text{rxn}_5, \text{rxn}_6, \text{rxn}_7, \text{rxn}_8]. \quad (23)$$

Using this gene pool to generate two children, results in mechanisms such as:

$$\mathbf{M}_{\text{child}_1} = [\text{rxn}_1, \text{rxn}_2, \text{rxn}_5, \text{rxn}_6]. \quad (24)$$

and

$$\mathbf{M}_{\text{child}_2} = [\text{rxn}_1, \text{rxn}_5, \text{rxn}_7, \text{rxn}_8]. \quad (25)$$

Crossover can be chosen to occur over specific reactions as described above, or in the reactant and/or product vectors separately. The latter will be advantageous when examining chemical processes with large number of chemical species.

Once the new generation of mechanisms is created using crossover, a random number of the newly-generated mechanisms are selected to be mutated. Mutation is not performed on the  $n_{\text{best}}$  elite mechanisms, only on the  $n_{\text{new}}$  new mechanisms generated using crossover. This is to retain the best solutions, otherwise the mutation could take an optimal solution and change it, removing important reaction information from the overall gene pool. Throughout this work we use a mutation rate  $\mathcal{M}$  of 0.1, meaning 10% of the new mechanisms are mutated. In the mutation procedure, first, a random mechanism is selected from  $n_{\text{new}}$  mechanisms. Next, a random reaction is selected from the selected mechanism and is substituted for another reaction from the original reaction list, with all reactions on this list being equally probable. Finally, the new mechanism with the substituted reac-



tion is checked to see if it satisfies the constraint that all chemical species in the dataset are included. If so, the new mechanism is substituted with original mechanism before mutation. If not, then the mutation procedure starts over by selecting a new mechanism to mutate, and the original mechanism stays on the list of mechanisms.

The rate constants in the new generation are then fit to construct the new generation of fit mechanisms  $\mathbf{G}_{i+1}^{(\text{fit})}$  and a corresponding set of values for the loss function  $\mathcal{L}_{i+1}$ . Then, the sort  $\rightarrow$  crossover  $\rightarrow$  mutate algorithm starts again until a set number of generations are evolved.

## 2.2.4 Final generation and mechanism selection

Because islanding is used in the SISR genetic algorithm, after the final generation of mechanisms is generated there is not a single best solution but instead a collection of best solutions, one for each island. Therefore, final selection of the overall best mechanism must be made. There are several ways to approach this problem. Here, we employ an approach based on multiobjective optimization. Specifically, we seek the mechanism that minimizes the discrepancy (the MSE) between the scaled ground truth concentration data and the data generated by the extracted mechanism:

$$\mathcal{L}_c \equiv \frac{1}{Nm} \sum_{j=1}^m \sum_{i=1}^N \left( \frac{[\hat{S}_i](t_j)}{\max([\hat{S}_i])} - \frac{[S_i](t_j)}{\max([S_i])} \right)^2, \quad (26)$$

and that also minimizes the complexity of the derived mechanism<sup>65</sup>. Complexity metrics in symbolic regression are used to quantify the simplicity or sparsity of the generated models. These metrics penalize overly complex expressions that do not significantly improve predictive performance<sup>55,66–68</sup>. Notice in Eq. (26) that while we discover the total mechanisms in the derivative space, the final determination of the best mechanism is made based on the concentration error, which is then coupled with a complexity metric.

The complexity metric applied in this work arises from counting the nodes in an expression tree that represents the kinetic equations for a mechanism (see Eq. (11)). Symbolic expression trees represent analytical functions in a hierarchical structure, where each node represents an operation or operand. The overall complexity metric is a sum over the number of nodes in the expression tree for each species in a mechanism:

$$\text{Complexity}(\mathbf{M}) = \sum_{i=1}^N |T_i|, \quad (27)$$

where  $T_i$  is the expression tree (the graph) of the kinetic equation for the  $i$ -th species and  $|T_i|$  is the number of nodes in that tree. Example expression trees are shown in Fig. 2.

Several modifications to typical expression tree construction are implemented:

1. We want larger stoichiometric coefficients to contribute more to the complexity in comparison to smaller coefficients. Therefore, all stoichiometric coefficients are written as a separate node or subtree. If the stoichiometric coefficient  $s = 1$ , it is written as a single node. If the stoichiometric coefficient

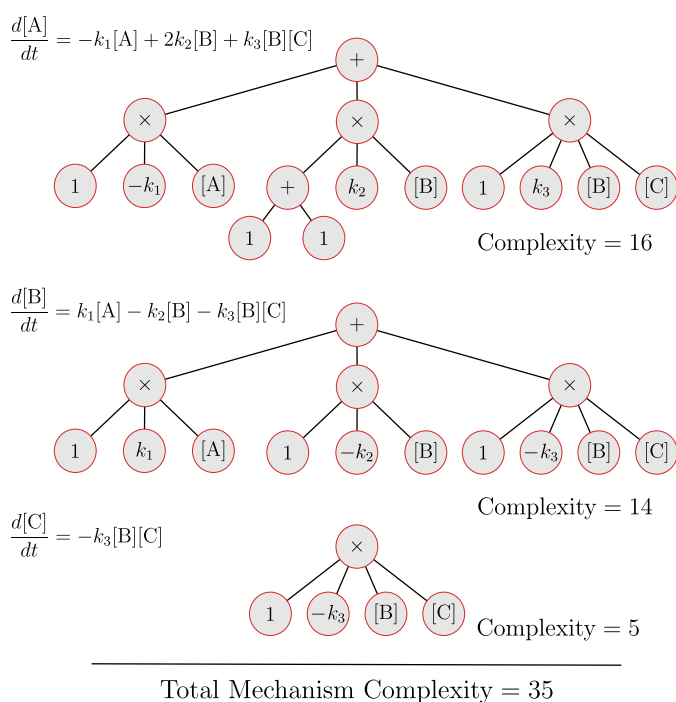


Fig. 2 Expression tree complexity analysis for the example mechanism shown in Eq. (28)

$s > 1$ , it is expressed as a subtree with  $s + 1$  nodes where the additional node is due to the “+” operation.

2. The “power” operation is not used to write nonlinear terms. Instead we write nonlinear terms such as  $[A]^2$  using the product operation. This is because terms like  $[A]^2$  should contribute the same complexity as terms like  $[A][B]$ .

In general, chemical reaction mechanisms have specific mathematical forms that contain polynomials but do not include other types of functions such as trigonometric or exponential functions. This simplifies the complexity calculation because we do not have to decide how to weight these different functions in the complexity hierarchy.

To illustrate how the complexity metric is implemented, consider the example mechanism



that is described by the set of kinetic equations

$$\begin{aligned} \frac{d[A]}{dt} &= -k_1[A] + 2k_2[B] + k_3[B][C] \\ \frac{d[B]}{dt} &= k_1[A] - k_2[B] - k_3[B][C] \\ \frac{d[C]}{dt} &= -k_3[B][C] \end{aligned} \quad (29)$$

Figure 2 illustrates the expression trees for each of the three equa-



tions in this system. Three key points are:

1. In the subtree for the term  $2k_2[B]$  in the  $\frac{d[A]}{dt}$  equation, note that the stoichiometric coefficient is broken into a sum. This adds a penalty on the stoichiometric values in the expression complexity.
2. Nonlinear terms such as  $k_1[A]^2$  contribute more complexity than linear terms such as  $k_1[A]$ . This agrees with physical intuition that nonlinear expressions are more complex than linear expressions.
3. All rate constants contribute the same to the complexity. Meaning no penalty is placed on the values of the rate constants. This is to avoid problems with coefficient thresholding that can arise in other methods.

The overall goal of the SISR procedure is to solve the multiobjective optimization problem

$$\min_{\mathbf{M}, \mathbf{k}} (\mathcal{L}_c, \text{Complexity}), \quad (30)$$

meaning we want to find the mechanism and the corresponding rate constant values that minimize the concentration error with respect to the ground truth data and also minimizes the complexity.

### 3 Results

#### 3.1 Sequential Linear Mechanism

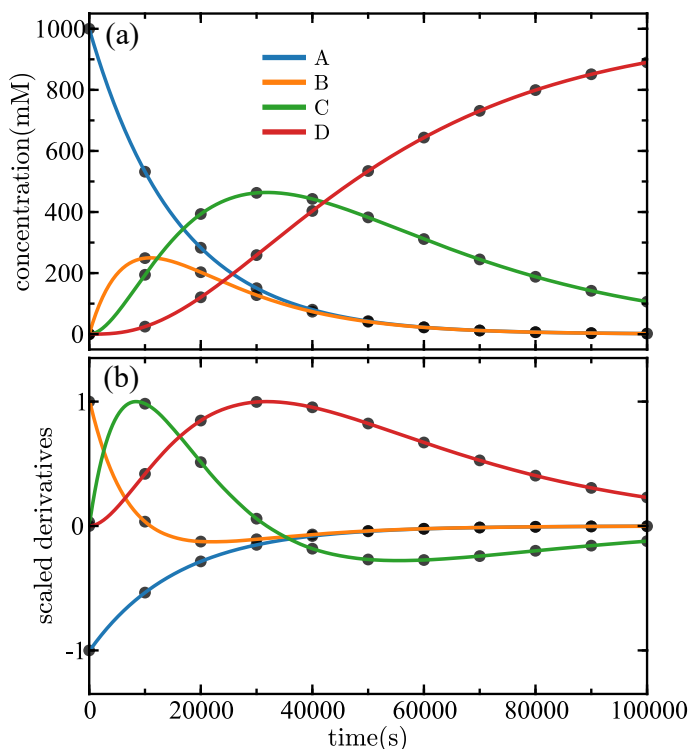
The first mechanism we examine is paradigmatic sequential linear mechanism



that is described by the set of kinetic equations for the time evolution of the concentrations:

$$\begin{aligned} \frac{d[A]}{dt} &= -k_1[A] \\ \frac{d[B]}{dt} &= k_1[A] - k_2[B] \\ \frac{d[C]}{dt} &= k_2[B] - k_3[C] \\ \frac{d[D]}{dt} &= k_3[C] \end{aligned} \quad (32)$$

with rate constants  $k_1 = 6.312 \times 10^{-5} \text{ s}^{-1}$ ,  $k_2 = 1.262 \times 10^{-4} \text{ s}^{-1}$ , and  $k_3 = 3.156 \times 10^{-5} \text{ s}^{-1}$  and concentrations given in millimolar. The search space involves 142 reactions that can be combined into  $\sim 1.1 \times 10^{10}$  possible mechanisms considering mechanisms with 2–6 reactions. So the search space is large. The kinetic equations were integrated using the Explicit Runge-Kutta method. We used this integration method for all examples in this manuscript. 500 equally spaced data points over the time interval [0, 100000] in units of seconds was fed into the SISR algorithm. For this mechanism, the SISR method was evolved for 10 generations over 5 islands with number of reactions  $|\mathbf{M}| = 2 - 6$ . A population



**Fig. 3** Time evolution of the (a) concentrations and (b) scaled derivatives—Eq. (34)—of each species in the sequential linear mechanism given in Eq. (32). The solid lines are the results of the SISR method and the corresponding black markers are a subset of the data used by SISR to extract the reaction mechanism and fit the rate constants. The concentrations are shown in units of millimolar and time is shown in units of seconds.

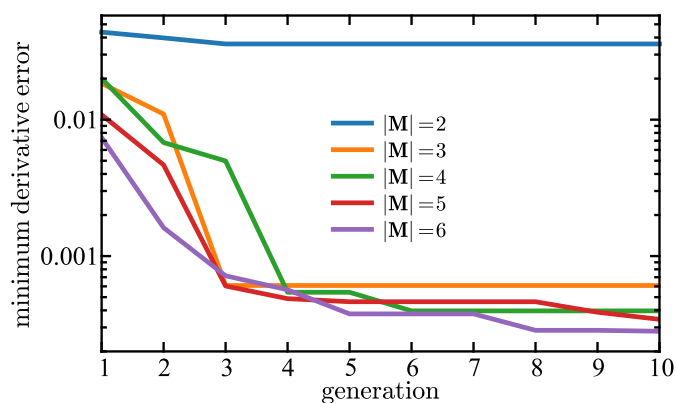
size of 2000 was used for the  $|\mathbf{M}| = 3 - 6$  islands while a population size of 500 was used for  $|\mathbf{M}| = 2$  island because of the smaller number of possible mechanisms on that island. The maximum reaction order was  $O = 2$ . The mutation rate was  $\mathcal{M} = 0.1$ . The elitism was  $\mathcal{E} = 0.1$  (meaning the top 10% of the solutions were retained across generations) except for the  $|\mathbf{M}| = 2$  island where the elitism was  $\mathcal{E} = 0.4$  due to the smaller population size.

The symbolic reaction extracted by SISR is:

$$\begin{aligned} \frac{d[\hat{A}]}{dt} &= -\hat{k}_1[\hat{A}] \\ \frac{d[\hat{B}]}{dt} &= \hat{k}_1[\hat{A}] - \hat{k}_2[\hat{B}] \\ \frac{d[\hat{C}]}{dt} &= \hat{k}_2[\hat{B}] - \hat{k}_3[\hat{C}] \\ \frac{d[\hat{D}]}{dt} &= \hat{k}_3[\hat{C}] \end{aligned} \quad (33)$$

which is exactly the true mechanism. The hat notation signifies that the rate constants and concentrations arise from the fits to the data. The rate constants in the extracted reaction mechanism are  $\hat{k}_1 = 6.310 \times 10^{-5} \text{ s}^{-1}$ ,  $\hat{k}_2 = 1.262 \times 10^{-4} \text{ s}^{-1}$ , and  $\hat{k}_3 = 3.156 \times 10^{-5} \text{ s}^{-1}$  in excellent agreement (0.0317% error or better) with the true rate constants. The results of the fitting are shown in Fig. 3. The concentrations predicted by the SISR method closely





**Fig. 4** Minimum derivative error,  $\min(\mathcal{L}_i)$ , as a function of generation using SISR to extract mechanisms from the ground truth data generated using the mechanism given in Eq. (32). The y-axis is shown on a log scale. Each curve is the result of a different island in the SISR method, where each island contains mechanisms with the number of reactions  $|\mathbf{M}|$  shown in the legend.

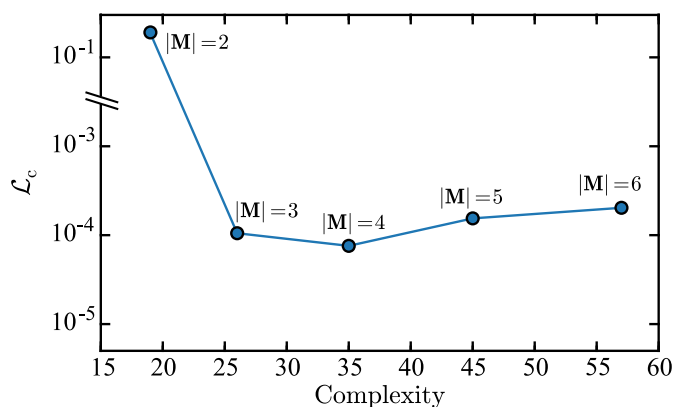
match the time evolution of the true mechanism, as illustrated in Fig. 3(a). In fact, at the presented level of visual fidelity, the ground truth data and the SISR result are indistinguishable. An important observation is that even though SISR searches over a symbolic space of nonlinear functions, the true linear mechanism is selected as the best model. Figure 3(b) illustrates a comparison between ground truth and SISR results for the scaled numerical derivatives,

$$\frac{\frac{d[S_i]}{dt}}{\max\left(\left|\frac{d[S_i]}{dt}\right|\right)}, \quad (34)$$

demonstrating remarkable quantitative alignment. Note that the rate constants for each total reaction mechanism and the MSE used to sort mechanisms on each island are calculated in the derivative space. So, while the overall goal is to develop a mechanism that results in time series concentration profiles, analyzing the SISR results in the derivative space is illustrative of results for the fitting and sorting procedure.

We have found that for the linear sequential mechanism, the SISR search converges to the true symbolic mechanism using as few as 20 data points in the fitting procedure. This supports the possibility that SISR will perform well on sparse kinetic data, for example, on the types of data that could be generated in some experimental setups. As expected, the more data points that are used in the fitting the closer the extracted rate constants become to the true rate constants. It is also interesting to note that the fitting method gives good agreement even when a relatively small number of data points are used.

Figure 4 illustrates how the minimum derivative error for each generation  $i$ , i.e.,  $\min(\mathcal{L}_i)$ , changes over 10 generations of the SISR genetic algorithm. In this case, the data that was fed into the SISR method was concentration data from 50 equally spaced time points over the time interval  $[0, 100000]$ . Each color curve in Fig. 4 represents the results for a different island size in the SISR algorithm. Figure 4 demonstrates that the genetic evolution of mechanisms results in monotonically decreasing error over



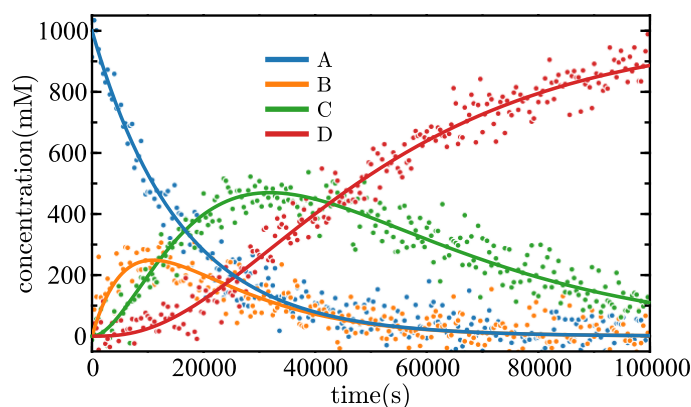
**Fig. 5** Complexity vs. concentration error  $\mathcal{L}_c$  for the sequential linear mechanism. Each marker corresponds to the best mechanism as calculated using the derivative error  $\mathcal{L}_{\text{der}}$  from the labeled island. The y-axis is shown on a log scale.

each generation. This shows that the SISR algorithm is optimizing the symbolic mechanism and the genetic search tends toward an optimal mechanism on each island. For  $|\mathbf{M}| = 2$ , only two reactions are included in each mechanism, and so the search space of possible mechanisms is small—approximately  $10^4$  mechanisms. Therefore, the optimal solution for that island is found after only a couple of generations because each generation includes a large portion of the total search space of possible mechanisms. However, the error generated on the  $|\mathbf{M}| = 2$  island is large compared to the other islands. The ground truth mechanism has  $|\mathbf{M}| = 3$  reactions, and there is a dramatic (approximately two orders of magnitude) drop in error when going from the  $|\mathbf{M}| = 2$  island to the  $|\mathbf{M}| = 3$  island. This illustrates an important result because to identify the correct mechanism a heuristic argument is to look for the optimal mechanism on the island that occurs immediately after a steep drop in error with respect to variation in mechanism size. Later we will give a more rigorous definition for mechanism selection, although we have found this heuristic observation to be an accurate identifier over the systems examined in this work.

In the derivative space, we see an interesting result which is that while the true mechanism has  $|\mathbf{M}| = 3$  reactions, the islands with  $|\mathbf{M}| = 4$ ,  $|\mathbf{M}| = 5$ , and  $|\mathbf{M}| = 6$  reactions actually generate lower error than the true mechanism, although this is not the case in the concentration space as we will show next. The cause of this reduction in error when adding spurious and erroneous reactions to the true mechanism size is that as more reactions are added, the algorithm has more coefficients to adjust in the fitting procedure, thereby fitting the data more closely. This means that the mechanisms on those islands are overfitting the data and doing so on symbolic functions that do not best fit the data. This also illustrates the need for a complexity measure when making the final mechanism selection, because if only minimum error was used then the wrong mechanism would be chosen in this case.

Figure 5 is a plot of the value for the loss function in the concentration space  $\mathcal{L}_c$  as a function of the mechanism complexity. Each marker in the plot represents the mechanism with the lowest derivative error on each island. This Pareto front plot illustrates that the  $|\mathbf{M}| = 3$  mechanism is the point on the Pareto front where





**Fig. 6** Time evolution of the concentrations of each species in the sequential linear mechanism given in Eq. (32). The SISR fit for each species extracted from noisy data is shown by a solid curve. The markers are a subset of the noisy data used by SISR.

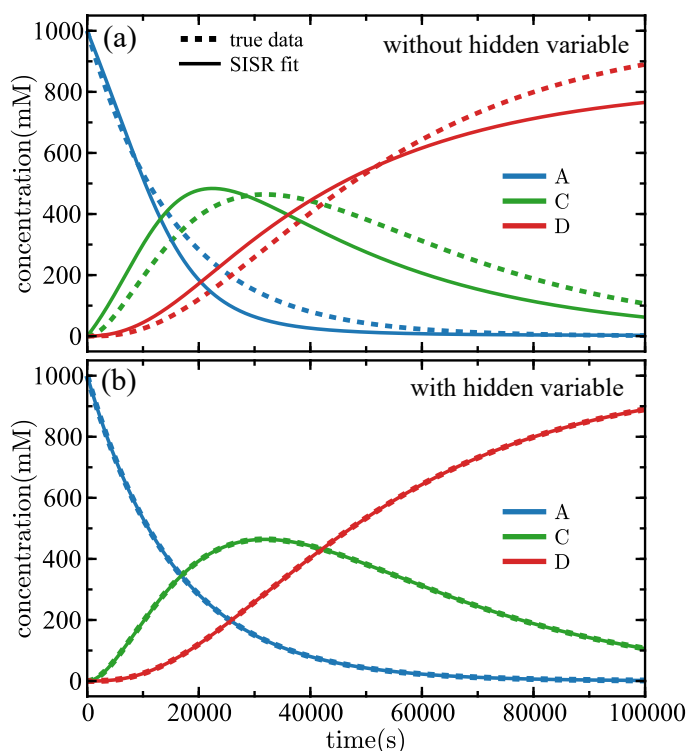
increasing complexity results in limited improvement in the accuracy of the mechanism. Therefore the  $|\mathbf{M}| = 3$  mechanism is the optimal solution. This can be observed because there is a steep drop in error when going from the  $|\mathbf{M}| = 2$  point to the  $|\mathbf{M}| = 3$  point, and while the complexity for  $|\mathbf{M}| = 2$  is smaller, the trade-off in concentration error is too dramatic to make it the optimal solution. When comparing the  $|\mathbf{M}| = 3$  to  $|\mathbf{M}| = 4$  mechanisms, there is limited improvement in concentration error when adding a new reaction, however there is a large increase in complexity. Interestingly, the  $|\mathbf{M}| = 5$  and  $|\mathbf{M}| = 6$  mechanisms result in an increase in error with respect to the  $|\mathbf{M}| = 3$  mechanism. This illustrates that including more terms in the symbolic mechanisms, i.e., including more reactions, does not necessarily result in improved accuracy but will generally result in a higher complexity. The error metrics used to construct the Pareto front in Fig. 5 are shown in Table 1 along with the corresponding derivative error for each mechanism.

**Table 1** SISR error metrics for the sequential linear mechanism on each island.

$ \mathbf{M} $	$\mathcal{L}_{\text{der}}$	$\mathcal{L}_{\text{c}}$	Complexity
2	$3.59 \times 10^{-2}$	$1.88 \times 10^{-1}$	19
3	$6.09 \times 10^{-4}$	$1.05 \times 10^{-4}$	26
4	$3.97 \times 10^{-4}$	$7.58 \times 10^{-5}$	35
5	$3.44 \times 10^{-4}$	$1.54 \times 10^{-4}$	45
6	$2.81 \times 10^{-4}$	$2.03 \times 10^{-4}$	57

### 3.1.1 Performance on Noisy Data

Noise can be a prominent feature in experiential and simulated chemical kinetics data, for example, in data obtained from molecular dynamics simulations using reactive force fields. Therefore, examining the robustness of the SISR approach in the presence of noise is an important metric. To this end, we examined the performance of SISR on noisy data by adding Gaussian noise sampled from the normal distribution  $\mathcal{N}(0,50)$  to the same dataset used in the deterministic solutions of the sequential linear mechanism given in Eq. (32). A Savitzky-Golay (SG) filter was applied to the concentration data, and that filtered data was used as the



**Fig. 7** Time evolution of the concentrations of the species in the sequential linear mechanism given in Eq. (32). Panel (a) shows the case in which data for species A, C, and D are fed into SISR but no hidden variable (a new chemical species) is allowed in the mechanism and panel (b) shows the same case but with a hidden variable being allowed. In both cases, the error function only includes species A, C, and D. The solid lines are the SISR fit and the dashed lines are the true data.

SISR input. This is a general approach that has worked previously due to known problems with numerical derivative calculations in the presence of noise<sup>64</sup>. The derivatives are computed after the smoothing is performed. The result of the SISR on the noisy data is shown in Fig. 6, with excellent agreement observed between the noisy data and the mechanism predicted by SISR. The SISR method is able to find correct mechanism (the same as in Eq.32). The rate constants extracted from the data were  $\hat{k}_1 = 6.346 \times 10^{-5} \text{ s}^{-1}$ ,  $\hat{k}_2 = 1.279 \times 10^{-4} \text{ s}^{-1}$ , and  $\hat{k}_3 = 3.091 \times 10^{-5} \text{ s}^{-1}$  in strong agreement with the true rate constants. Note the specific rate constant values extracted from noisy data will depend on the specific realization of the noise. An important point is that the SISR method is stochastic and not guaranteed to find the optimal solution (in this case meaning the correct mechanism) for each realization of the noise. Overall, this result illustrates the ability of SISR to extract the true reaction mechanism on data with high levels of noise.

### 3.1.2 Hidden Variables and Intermediates

The SISR workflow can also be used to detect the presence of hidden variables (such as unknown chemical intermediates) in a set of chemical concentration data. The specific question we want to address is: Given a set of concentration data, does including more chemical species (like intermediates) in the symbolic regression part of the SISR algorithm beyond what is present in



the data result in a more accurate reaction mechanism? To address this question, we consider the case of the sequential linear mechanism but where only a subset of the data is fed into SISR, for example concentration data for species A, B, and D or species A, C, and D. The SISR method is then used to detect the presence of the missing species (the hidden intermediate species). In all cases considered in this section, the loss function is computed by only using species with available concentration data—the hidden intermediate is excluded from the error calculation.

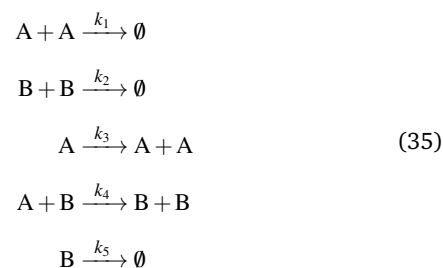
Figure 7(a) is the SISR result for the case in which data from species A, C, and D are used (the unknown intermediate in this case is species B), and SISR is applied to find a mechanism containing only those species. In this case, a poor fit and mechanism is obtained for the available data. The reason for this is that, without including the hidden intermediate, the model cannot reproduce the correct time evolution of the concentrations or reaction pathways that connect the observed species. Now, compare those results with the results in Figure 7(b) where SISR is allowed to search for mechanisms that include one additional (previously hidden) species. In this case, the recovered mechanism correctly identifies the presence of the missing intermediate and yields a significantly improved fit to the data. Specifically, the error computed using Eq. 15 drops by a factor of approximately  $10^5$  compared to the case shown in Figure 7(a) that does not include an intermediate in the reaction mechanism. It is interesting to note that even though only data from three species is used, the SISR approach converges to the exact correct underlying mechanism. Therefore, including the intermediate species in the reaction mechanism, despite the absence of direct data for that species, yields a significantly improved mechanism. This demonstrates that the dataset implicitly contains evidence of a hidden intermediate, which is consistent with the known (see Eq. (32)) underlying reaction mechanism.

We have also confirmed that a similar level of improvement is observed using hidden variables in the reaction mechanism when other species are removed from the data set. So, for example, when data for species A, B, and D are used with C being excluded and when data from species A, B, and C are used with D being excluded. In all cases we have studied for this mechanism, the application of SISR yields the exact true mechanism and also detects the presence of a hidden variable.

This proof-of-concept illustrates how SISR can be used to detect and then fit hidden chemical intermediates. Further work in this area will focus on constructing a multidimensional Pareto front that also includes the number of species involved in the process as an optimization dimension. In a many scenarios, for example in many biological systems, in combustion modeling, and in microkinetic modeling, there could be a large number of hidden variables. The presented results detail the case when there is only a single hidden variable, and therefore opportunities exist for future work to examine the ability of SISR to handle more complex cases with multiple hidden variables.

### 3.2 Lotka-Volterra with Social Friction

Next, we applied SISR to the Lotka-Volterra with Social Friction mechanism examined in Ref. 41 using the Reactive SINDy method. This mechanism exhibits the types of oscillatory behaviors seen in some biochemical systems<sup>69</sup>, for example in some viruses and in susceptible cells. The specific mechanism is



and involves two species A and B. Where  $\emptyset$  denotes the annihilation or irreversible removal of reactants. The corresponding kinetic equations for the system are

$$\begin{aligned} \frac{d[A]}{dt} &= -2k_1[A]^2 + k_3[A] - k_4[A][B] \\ \frac{d[B]}{dt} &= -2k_2[B]^2 + k_4[A][B] - k_5[B] \end{aligned} \quad (36)$$

with rate constant values given by  $k_1 = 0.1$ ,  $k_2 = 0.1$ ,  $k_3 = 1$ ,  $k_4 = 1$ ,  $k_5 = 1$ . The reaction forms in Eq. (35) can be interpreted as effective representations of an underlying atomic mechanism (see for example, Ref. 70). The individual reactions used are representative of, for example: autocatalytic reactions, branching reactions, and interactions with sinks, among other chemical processes. In this mechanism, the stoichiometric coefficients represent the conversion of reactants to products in the effective representation.

To generate this mechanism, a population of mechanisms was evolved for 20 generations over 6 islands with  $|\mathbf{M}| = 2 - 7$  reactions. The population size was 2000 for the islands with  $|\mathbf{M}| = 3 - 7$  reactions and 500 for the  $|\mathbf{M}| = 2$  island. The mutation rate was 0.1 and the elitism was 0.1 for all islands except the  $|\mathbf{M}| = 2$  island where the elitism was 0.4 due to the smaller population size. The maximum reaction order was  $O = 2$ . The ground truth data was generated using 1000 equally-space points over the time interval  $[0, 20]$ . The search space involves 57 reactions that can be combined into  $\sim 3.1 \times 10^8$  possible mechanisms considering mechanisms with  $2 - 7$  reactions.

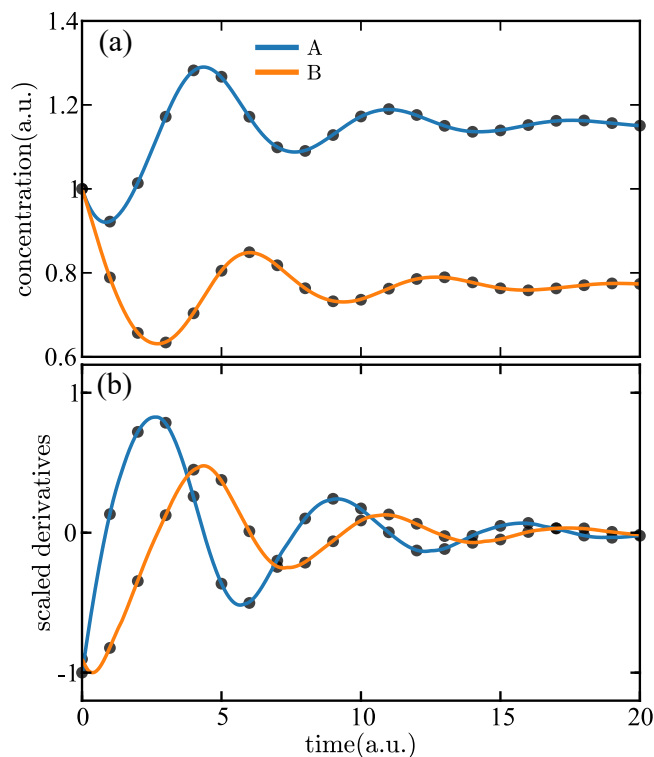
The model extracted from the data by the SISR method is

$$\begin{aligned} \frac{d[\hat{A}]}{dt} &= -2\hat{k}_1[\hat{A}]^2 + \hat{k}_3[\hat{A}] - \hat{k}_4[\hat{A}][\hat{B}] \\ \frac{d[\hat{B}]}{dt} &= -2\hat{k}_2[\hat{B}]^2 + \hat{k}_4[\hat{A}][\hat{B}] - \hat{k}_5[\hat{B}] \end{aligned} \quad (37)$$

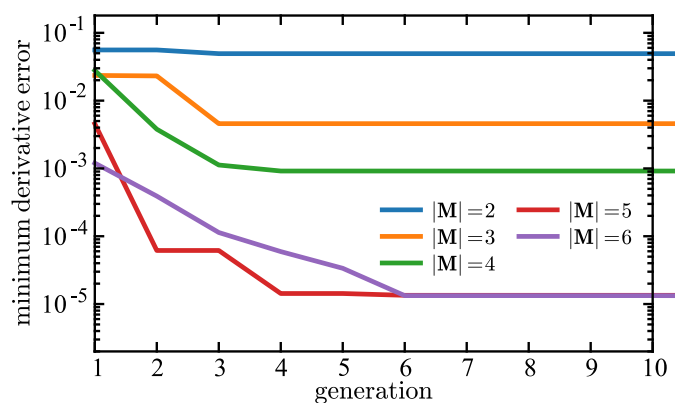
which is, again, in exact agreement with the symbolic form for the ground truth reaction mechanism. The values of the rate constants of this total mechanism were  $\hat{k}_1 = 0.1003$ ,  $\hat{k}_2 = 0.1029$ ,  $\hat{k}_3 = 1.001$ ,  $\hat{k}_4 = 1.001$ ,  $\hat{k}_5 = 0.996$ , in strong agreement (2.81% error or better) with the true rate constants.

The SISR results are shown in Fig.8 with panel (a) containing a comparison between the true and predicted concentration





**Fig. 8** Time evolution of the (a) concentrations and (b) scaled derivatives—Eq. (35)—of each species in the Lotka-Volterra mechanism with social friction given in Eq. (36). The solid lines are the results of the SISR method and the corresponding black markers are a subset of the data used by SISR to extract the reaction mechanism and fit the rate constants. The concentrations and time are shown in arbitrary units (a.u.).



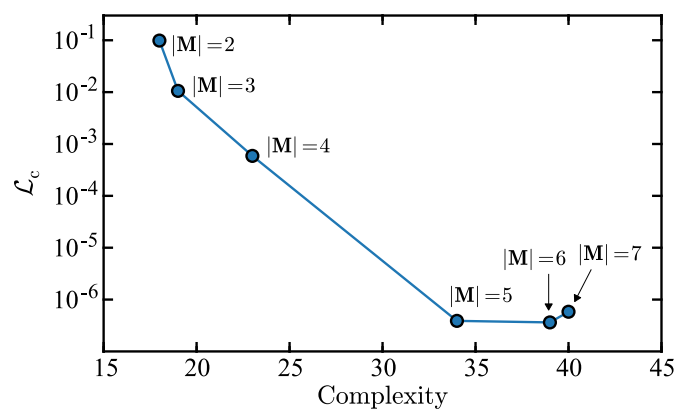
**Fig. 9** Minimum derivative error,  $\min(\mathcal{L}_i)$ , as a function of generation using SISR to extract mechanisms from the ground truth data generated using the Lotka-Volterra mechanism given in Eq. (35). Each curve is the result of a different island in the SISR method, where each island contains mechanisms with the number of reactions  $|\mathbf{M}|$  shown in the legend.

data and panel (b) containing the same comparison for the scaled derivative data. Excellent agreement is observed between the SISR result and the ground-truth data in both cases. The symbolic reaction mechanism is fitted on the derivative data, and Fig.8 (b) illustrates a primary advantage in using SISR: due to the inclusion of stoichiometric constraints, there is a distinct lack of overfitting the discovered mechanism, i.e., in the dynamical system that is extracted from the data.

Figure 9 illustrates how the minimum derivative error changes over 10 generations when applying SISR to the Lotka-Volterra mechanism. Again, as in the previous mechanism, the genetic evolution results in monotonically decreasing error over each generation for every island size. The error generated on the  $|\mathbf{M}| = 2$  island is the largest. As more reactions are added to the mechanism, the error decreases, which can be observed by comparing the results for each island size. The ground truth mechanism has  $|\mathbf{M}| = 5$  reactions, and the most dramatic (approximately two orders of magnitude) drop in error is observed when going from the  $|\mathbf{M}| = 4$  island to the  $|\mathbf{M}| = 5$  island. So for this reaction mechanism, the heuristic argument that the optimal mechanism occurs on the island that follows immediately after a steep drop in error with respect to variation in mechanism size would yield the selection of the true mechanism. While the true mechanism has  $|\mathbf{M}| = 5$  reactions, the island with  $|\mathbf{M}| = 6$  reactions generates a lower error than  $|\mathbf{M}| = 5$  island, although this difference is small ( $1.35 \times 10^{-5}$  for  $|\mathbf{M}| = 5$  compared to  $1.33 \times 10^{-5}$  for  $|\mathbf{M}| = 6$ ). The reduction in error when adding spurious reactions to the mechanism is due to the algorithm having more coefficients to fit in the symbolic model.

Figure 10 is a plot of the value for the loss function in the concentration space  $\mathcal{L}_c$  as a function of the mechanism complexity for the Lotka-Volterra mechanism. The Pareto front shown in the plot illustrates that the  $|\mathbf{M}| = 5$  mechanism is the point where increasing complexity by adding more reactions results in limited improvement in the accuracy of the mechanism. Therefore the  $|\mathbf{M}| = 5$  mechanism was chosen as the optimal solution. There is a steep drop in error when going from the  $|\mathbf{M}| = 2$  point to the





**Fig. 10** Complexity vs. concentration error  $\mathcal{L}_c$  for the Lotka-Volterra mechanism. Each marker corresponds to the best mechanism as calculated using the derivative error  $\mathcal{L}_{der}$  from the labeled island. The y-axis is shown on a log scale.

$|\mathbf{M}| = 3$  point and from the  $|\mathbf{M}| = 3$  point to the  $|\mathbf{M}| = 4$ . Comparing the  $|\mathbf{M}| = 4$  point and the  $|\mathbf{M}| = 5$  point (the true mechanism size) we see approximately a three orders of magnitude decrease in error. Therefore, despite the smaller mechanisms having less complexity, the tradeoff in increased error is too dramatic to make the smaller mechanisms the optimal solution. Compare this with the results for the  $|\mathbf{M}| = 5$  and  $|\mathbf{M}| = 6$  mechanisms where there is limited improvement in concentration error when adding a new reaction, but there is a significant increase in complexity. The  $|\mathbf{M}| = 7$  mechanism shows an increase in error with respect to the  $|\mathbf{M}| = 5$  and  $|\mathbf{M}| = 6$  mechanisms. The error metrics used to construct the Pareto front in Fig. 10 are shown in Table 2 along with the corresponding derivative error for each mechanism.

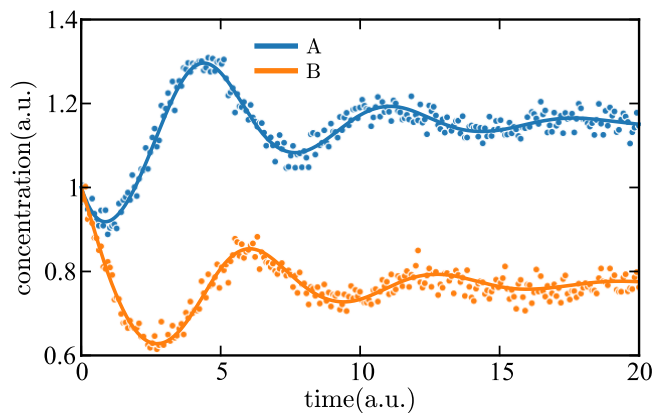
**Table 2** SISr error metrics for the Lotka-Volterra mechanism on each island.

$ \mathbf{M} $	$\mathcal{L}_{der}$	$\mathcal{L}_c$	Complexity
2	$4.93 \times 10^{-2}$	$9.86 \times 10^{-2}$	18
3	$4.58 \times 10^{-3}$	$1.06 \times 10^{-2}$	19
4	$9.16 \times 10^{-4}$	$5.89 \times 10^{-4}$	23
5	$1.35 \times 10^{-5}$	$3.88 \times 10^{-7}$	34
6	$1.33 \times 10^{-5}$	$3.62 \times 10^{-7}$	39
7	$1.32 \times 10^{-5}$	$5.82 \times 10^{-7}$	40

### 3.2.1 Comparison to SINDy

To illustrate how the SISr method compares to the well-used SINDy approach, we applied SINDy to the Lotka-Volterra mechanism. Applying SINDy using the SR3 (sparse relaxed regularized regression) method<sup>64,71</sup> with a threshold value of 0.01 resulted in

$$\begin{aligned} \frac{d[\hat{A}]}{dt} &= 1.030[\hat{A}] - 0.113[\hat{B}] - 0.227[\hat{A}]^2 \\ &\quad + 0.050[\hat{B}]^2 - 0.934[\hat{A}][\hat{B}] \\ \frac{d[\hat{B}]}{dt} &= -0.224[\hat{A}] - 0.630[\hat{B}] + 0.065[\hat{A}]^2 \\ &\quad - 0.501[\hat{B}]^2 + 1.073[\hat{A}][\hat{B}] \end{aligned} \quad (38)$$



**Fig. 11** Time evolution of the concentrations of each species in the Lotka-Volterra mechanism given in Eq. (36). The SISr fit for each species extracted from noisy data is shown by a solid curve. The markers are a subset of the noisy data used in the SISr process.

which illustrates a principal problem that can arise when using the SINDy method without any physical constraints—overfitting in the derivative space. Notice that every possible term in the symbolic dynamical system up to second-order has a non-zero coefficient and that, correspondingly, the discovered system is not in agreement with the true mechanism. Increasing the SINDy threshold to a value of 0.25 in order to include fewer terms in the mechanism, i.e., to promote SINDy producing a sparser reaction mechanism, resulted in

$$\begin{aligned} \frac{d[\hat{A}]}{dt} &= 0.707[\hat{A}] - 0.918[\hat{A}][\hat{B}] \\ \frac{d[\hat{B}]}{dt} &= -0.951[\hat{B}] - 0.239[\hat{B}]^2 + 0.983[\hat{A}][\hat{B}] \end{aligned} \quad (39)$$

which is again not in agreement with the true mechanism.

Applying the Reactive SINDy method will result in similar problems to regular SINDy unless hyperparameters are tuned for the specific mechanism and a suitable basis set of candidate reactions are constructed. Although it should be noted that after performing these tuning and construction tasks, Reactive SINDy generates an overall mechanism (reactions and rate constants) in excellent agreement with the true Lotka-Volterra mechanism<sup>41</sup>.

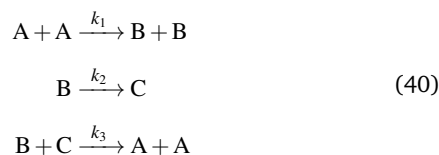
### 3.2.2 Performance on Noisy Data

We also examined the performance of SISr on the Lotka-Volterra mechanism with noisy concentration data. The results are shown in Fig. 11. To generate the noisy data, Gaussian noise sampled from the normal distribution  $\mathcal{N}(0, 0.02)$  was added to the deterministic solutions of the system in Eq. (36). We then applied a SG filter to the concentration data. The result of using this data as input to SISr is shown in Fig. 11, with excellent agreement observed between the noisy data and the mechanism predicted by SISr. This illustrates the ability of SISr to perform on concentration data with noise for an oscillatory reaction, a common situation in biological systems.



### 3.3 Nonlinear Mechanism with Fast/Slow Dynamics

In order to test a nonlinear mechanism with fast/slow dynamics, i.e., a mechanism that has processes that evolve over disparate timescales, we examine the mechanism



The kinetic equations for this system are

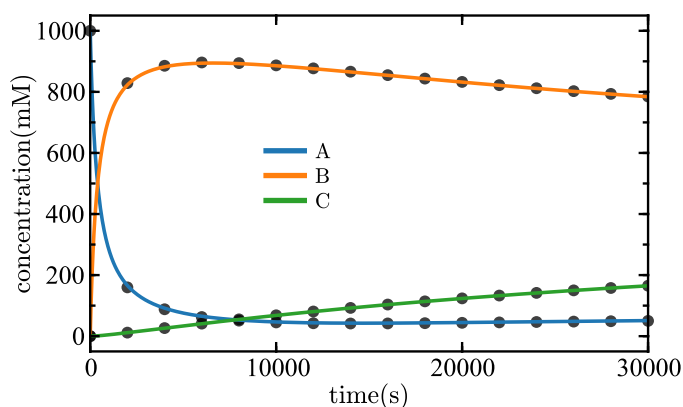
$$\begin{aligned} \frac{d[A]}{dt} &= -2k_1[A]^2 + 2k_3[B][C] \\ \frac{d[B]}{dt} &= 2k_1[A]^2 - k_2[B] - k_3[B][C] \\ \frac{d[C]}{dt} &= k_2[B] - k_3[B][C] \end{aligned} \quad (41)$$

with rate constants  $k_1 = 1.319 \times 10^{-6} \text{ mM}^{-1} \text{ s}^{-1}$ ,  $k_2 = 9.125 \times 10^{-6} \text{ s}^{-1}$ , and  $k_3 = 2.756 \times 10^{-8} \text{ mM}^{-1} \text{ s}^{-1}$  and concentrations being given in millimolar. The fast/slow dynamics are obtained because the numerical value of the reaction rate  $k_3$  is orders of magnitude different than the other rate constants. In this mechanism, there is a rapid decline of [A] and the corresponding rapid increase in [B] (the fast processes) followed by a slow increase in [C] (the slow process), as shown in Fig. 12.

The data fed into SISR was 2000 equally spaced time points over the interval [0,20000] in units of seconds. The density of data was the same in the fast region as in the slow region. For this mechanism, the search space involves 54 reactions that can be combined into  $\sim 2.9 \times 10^7$  mechanisms. Respective population sizes of 2000 and 500 were evolved for 20 generations using islands with  $|\mathbf{M}| = 3 - 5$  and  $|\mathbf{M}| = 2$  reactions. The mutation rate was  $\mathcal{M} = 0.1$  and the elitism was  $\mathcal{E} = 0.1$ . The maximum reaction order was  $O = 2$ . The best fit model found by the SISR method is

$$\begin{aligned} \frac{d[\hat{A}]}{dt} &= -2\hat{k}_1[\hat{A}]^2 + 2\hat{k}_3[\hat{B}][\hat{C}] \\ \frac{d[\hat{B}]}{dt} &= 2\hat{k}_1[\hat{A}]^2 - \hat{k}_2[\hat{B}] - \hat{k}_3[\hat{B}][\hat{C}] \\ \frac{d[\hat{C}]}{dt} &= \hat{k}_2[\hat{B}] - \hat{k}_3[\hat{B}][\hat{C}] \end{aligned} \quad (42)$$

Again, in exact agreement with the true mechanism. This mechanism was selected from the Pareto front in the Complexity vs.  $\mathcal{L}_c$  space using the same multiobjective procedure described previously. The rate constants in the found reaction mechanism are  $\hat{k}_1 = 1.261 \times 10^{-6} \text{ mM}^{-1} \text{ s}^{-1}$ ,  $\hat{k}_2 = 9.121 \times 10^{-6} \text{ s}^{-1}$ , and  $\hat{k}_3 = 2.750 \times 10^{-8} \text{ mM}^{-1} \text{ s}^{-1}$  which very closely match the true rate constants (4.39% error or better). The results of the SISR model are shown in Fig. 12 with excellent agreement observed between the concentrations predicted by the SISR method and the concentrations calculated using the true mechanism. For times above 20000, the SISR model is forecasting, meaning no training data was used from that time interval. Excellent agreement is observed in the forecasted region.



**Fig. 12** Time evolution of the concentrations of each species in the nonlinear mechanism given in Eq. (41). The solid lines are the results of the SISR method and the corresponding black markers are a subset of the data used by SISR to extract the reaction mechanism and fit the rate constants. The concentrations are shown in units of millimolar and time is shown in units of seconds.

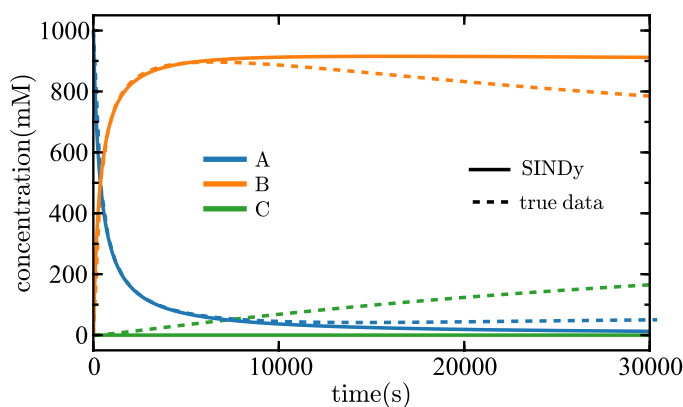
#### 3.3.1 Comparison to SINDy

To illustrate how the SISR method compares to SINDy, we applied SINDy using SR3 optimization with a threshold of  $10^{-8}$ . The data that was fed into SINDy was from 5000 equally spaced time points over the interval [0,50000]. This resulted in a set of kinetic equations:

$$\begin{aligned} \frac{d[\hat{A}]}{dt} &= -2.64 \times 10^{-6}[\hat{A}]^2 + 5.52 \times 10^{-8}[\hat{B}][\hat{C}] \\ \frac{d[\hat{B}]}{dt} &= 2.66 \times 10^{-6}[\hat{A}]^2 - 6.66 \times 10^{-8}[\hat{A}][\hat{B}] \\ &\quad + 1.27 \times 10^{-6}[\hat{A}][\hat{C}] - 1.49 \times 10^{-7}[\hat{B}][\hat{C}] \\ \frac{d[\hat{C}]}{dt} &= 3.01 \times 10^{-8}[\hat{B}][\hat{C}] \end{aligned} \quad (43)$$

which illustrates several problems: (a) overfitting in the derivative space, (b) a lack of stoichiometric information, and (c) the misidentification of chemical processes. To illustrate problem (c), note that the true mechanism has a process in the  $\frac{d[B]}{dt}$  and  $\frac{d[C]}{dt}$  equations involving a first-order reaction in species B, but the SINDy result does not. Figure 13 illustrates a comparison between the true data and the dynamical system derived using SINDy. Notice that the slow rise of species C is not captured in the SINDy model. Compare this with the SISR result in Fig. 12 where that slow process is well captured. It should be noted that SINDy was not developed for chemical reaction mechanisms and is overall agnostic to the specific physics of a system. Although this can be a strength in certain situations, it hinders the discovery of accurate chemical reaction mechanisms. The SISR method, however, encodes physical information about chemical reactions and stoichiometry and therefore performs better with respect to accuracy on the selected examples in this work.

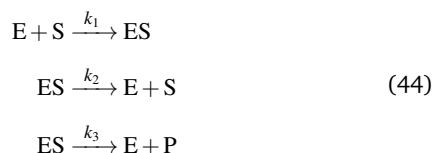




**Fig. 13** Time evolution of the concentrations of each species in the nonlinear mechanism given in Eq. (41). The solid lines are the results of the SINDy method and the dashed lines are the true data. The concentrations are shown in units of millimolar and time is shown in units of seconds.

### 3.4 Michaelis-Menten Kinetics

The final example we examine is the Michaelis-Menten (MM) kinetic model, an important mechanism in biochemistry<sup>72,73</sup>. The specific MM model we use is the traditional form



with the corresponding set of kinetic equations

$$\begin{aligned} \frac{d[\text{E}]}{dt} &= -k_1[\text{E}][\text{S}] + k_2[\text{ES}] + k_3[\text{ES}] \\ \frac{d[\text{S}]}{dt} &= -k_1[\text{E}][\text{S}] + k_2[\text{ES}] \\ \frac{d[\text{ES}]}{dt} &= k_1[\text{E}][\text{S}] - k_2[\text{ES}] - k_3[\text{ES}] \\ \frac{d[\text{P}]}{dt} &= k_3[\text{ES}] \end{aligned} \quad (45)$$

where E is an enzyme, S is a substrate, ES is an enzyme-substrate complex, and P is a product. The rate constants used are  $k_1 = 1 \text{ mM}^{-1}\text{s}^{-1}$ ,  $k_2 = 0.1 \text{ s}^{-1}$ , and  $k_3 = 1 \text{ s}^{-1}$ . The concentrations are given in millimolar and time is given in seconds.

The data fed into SISR was 400 equally spaced time points over the interval  $[0, 2]$  in units of seconds. For this mechanism, the search space involves 182 reactions. A population size of 4000 was evolved for 20 generations for islands with  $|\mathbf{M}| = 3 - 5$  reactions and a population size of 500 was used for the  $|\mathbf{M}| = 2$  island. The mutation rate was  $\mathcal{M} = 0.1$  and the elitism was  $\mathcal{E} = 0.1$ . The population size for the  $|\mathbf{M}| = 3 - 5$  islands was doubled from previous examples due to the larger number of possible reactions.

The maximum reaction order was  $O = 2$ . The SISR result is

$$\begin{aligned} \frac{d[\hat{\text{E}}]}{dt} &= -\hat{k}_1[\hat{\text{E}}][\hat{\text{S}}] + \hat{k}_2[\hat{\text{ES}}] + \hat{k}_3[\hat{\text{ES}}] \\ \frac{d[\hat{\text{S}}]}{dt} &= -\hat{k}_1[\hat{\text{E}}][\hat{\text{S}}] + \hat{k}_2[\hat{\text{ES}}] \\ \frac{d[\hat{\text{ES}}]}{dt} &= \hat{k}_1[\hat{\text{E}}][\hat{\text{S}}] - \hat{k}_2[\hat{\text{ES}}] - \hat{k}_3[\hat{\text{ES}}] \\ \frac{d[\hat{\text{P}}]}{dt} &= \hat{k}_3[\hat{\text{ES}}] \end{aligned} \quad (46)$$

which agrees exactly with the true mechanism. The SISR mechanism was selected using the previously described procedure by comparing the error and complexity of the best performing mechanism from each island and making a determination based on the shape of the Pareto front, specifically looking for the steepest drop in error which occurred between the  $|\mathbf{M}| = 2$  and  $|\mathbf{M}| = 3$  islands. Note that the  $|\mathbf{M}| = 4$  island produced the lowest concentration error, but, as in the previous examined mechanisms, comparing the complexity of the mechanisms and error in a multiobjective picture results in the  $|\mathbf{M}| = 3$  mechanism being chosen as the best overall mechanism. The fit rate constants for the extracted mechanism are  $\hat{k}_1 = 0.993 \text{ mM}^{-1}\text{s}^{-1}$ ,  $\hat{k}_2 = 0.084 \text{ s}^{-1}$ , and  $\hat{k}_3 = 1.000 \text{ s}^{-1}$ , compared to the true rate constants.

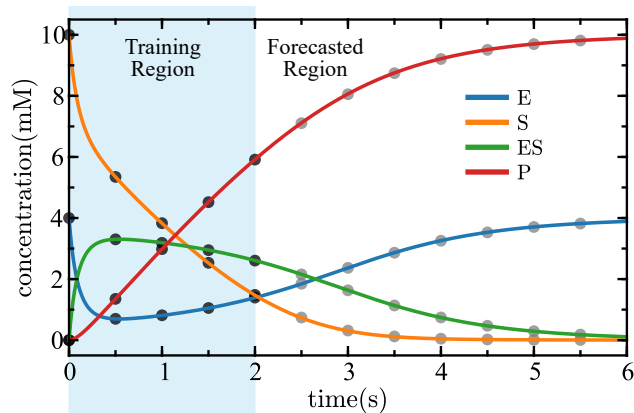
The results of SISR on the MM model is shown in Fig. 14, with excellent agreement observed between the SISR-predicted concentrations and the ground truth data. The dynamics of the MM model are complex and involve competing processes that occur over different time scales. There are fast processes such as the fast decay of species E and the corresponding fast rise of species ES, and also slow processes such as the rise of species E and P to the steady-state values. SISR captures these processes well.

The region shown in blue in Fig. 14 is the training region—data from this region was used to extract the mechanism and fit the rate constants. The white region in Fig. 14 shows the SISR result on data that was not used for mechanism discovery. Data in this region can be used to validate the SISR mechanism on unseen data and to assess the capability of SISR for time-series forecasting of chemical concentrations. Excellent agreement is observed between the ground truth data and the forecasted concentrations. This illustrates one of the principal advantages of using SISR for reaction discovery—because of the stoichiometrically-informed construction of the reaction mechanism, it is able to accurately forecast concentrations at future time points while avoiding the overfitting and extrapolation errors that are common in black-box machine learning.

### 3.5 Glucose Oxidation

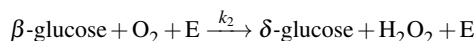
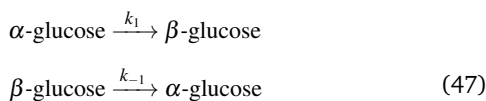
The final reaction mechanism we examine with SISR is glucose oxidation (GO), an important biochemical process. The specific GO mechanism we consider consists of the conversion from  $\alpha$ -glucose to  $\beta$ -glucose, the reverse reaction involving  $\beta$ -glucose to  $\alpha$ -glucose conversion, and the glucose oxidase (E) catalyzed reaction of  $\beta$ -glucose with oxygen to produce another glucose molecule— $\delta$ -glucose—and  $\text{H}_2\text{O}_2$ . The individual reactions for





**Fig. 14** Time evolution of the concentrations of each species in the Michaelis-Menten kinetic mechanism given in Eq. (45). The solid lines are the results of the SISR method and the corresponding black markers are a subset of the training data used by SISR to extract the reaction mechanism and fit the rate constants and the gray markers are testing data that was not used to fit the SISR model. Data from the light blue region is used to train SISR while data from the white region is forecasted. The concentrations are shown in units of millimolar and time is shown in units of seconds.

this model are



which give rise to the corresponding set of kinetic equations

$$\begin{aligned} \frac{d[\alpha]}{dt} &= -k_1[\alpha] + k_{-1}[\beta] \\ \frac{d[\beta]}{dt} &= k_1[\alpha] - k_{-1}[\beta] - k_2[\beta][\text{O}_2] \\ \frac{d[\delta]}{dt} &= k_2[\beta][\text{O}_2][\text{E}] \\ \frac{d[\text{E}]}{dt} &= 0 \\ \frac{d[\text{O}_2]}{dt} &= -k_2[\beta][\text{O}_2][\text{E}] \\ \frac{d[\text{H}_2\text{O}_2]}{dt} &= k_2[\beta][\text{O}_2][\text{E}] \end{aligned} \quad (48)$$

We assume that the enzyme has a steady state concentration of  $1\mu\text{M}$  and that we have *a priori* knowledge of the enzyme interactions. Two sets of rates constants are examined. The values in the first set are  $k_1 = 4.45 \times 10^{-4}\text{s}^{-1}$ ,  $k_{-1} = 3.03 \times 10^{-4}\text{s}^{-1}$ , and  $k_2 = 2.78 \times 10^{-3}\mu\text{M}^{-2}\text{s}^{-1}$ . In the second set, the value of  $k_2$  is increased by an order of magnitude to  $k_2 = 2.78 \times 10^{-2}\mu\text{M}^{-2}\text{s}^{-1}$ . These values are based on work in Ref. 74.

For the first set of rate constants, using SISR with a population size of 2000 and with 500 data points over the interval  $[0, 10000]$ ,

SISR discovered the correct reactions and reaction mechanism:

$$\begin{aligned} \frac{d[\hat{\alpha}]}{dt} &= -\hat{k}_1[\hat{\alpha}] + \hat{k}_{-1}[\hat{\beta}] \\ \frac{d[\hat{\beta}]}{dt} &= \hat{k}_1[\hat{\alpha}] - \hat{k}_{-1}[\hat{\beta}] - \hat{k}_2[\hat{\beta}][\hat{\text{O}}_2] \\ \frac{d[\hat{\delta}]}{dt} &= \hat{k}_2[\hat{\beta}][\hat{\text{O}}_2][\hat{\text{E}}] \\ \frac{d[\hat{\text{E}}]}{dt} &= 0 \\ \frac{d[\hat{\text{O}}_2]}{dt} &= -\hat{k}_2[\hat{\beta}][\hat{\text{O}}_2][\hat{\text{E}}] \\ \frac{d[\hat{\text{H}}_2\hat{\text{O}}_2]}{dt} &= \hat{k}_2[\hat{\beta}][\hat{\text{O}}_2][\hat{\text{E}}] \end{aligned} \quad (49)$$

The extracted rate constants were  $\hat{k}_1 = 4.45 \times 10^{-4}\text{s}^{-1}$ ,  $\hat{k}_{-1} = 3.23 \times 10^{-4}\text{s}^{-1}$ , and  $\hat{k}_2 = 2.73 \times 10^{-3}\mu\text{M}^{-2}\text{s}^{-1}$  in strong agreement with the true values. A comparison between input data and the SISR fit is shown in Fig. 15(a)

For the second set of rate constants, a population size of 4000 with 1000 data points over the time interval  $[0, 20000]$  in units of seconds was used. Again, the exact reaction mechanism was found using SISR, illustrating the ability of SISR to account for fast/slow dynamics. The extracted rate constants were  $\hat{k}_1 = 4.40 \times 10^{-4}\text{s}^{-1}$ ,  $\hat{k}_{-1} = 2.56 \times 10^{-4}\text{s}^{-1}$ , and  $\hat{k}_2 = 2.63 \times 10^{-2}\mu\text{M}^{-2}\text{s}^{-1}$  in good agreement with the true values. The result of the SISR fit is shown in Fig. 15(b).

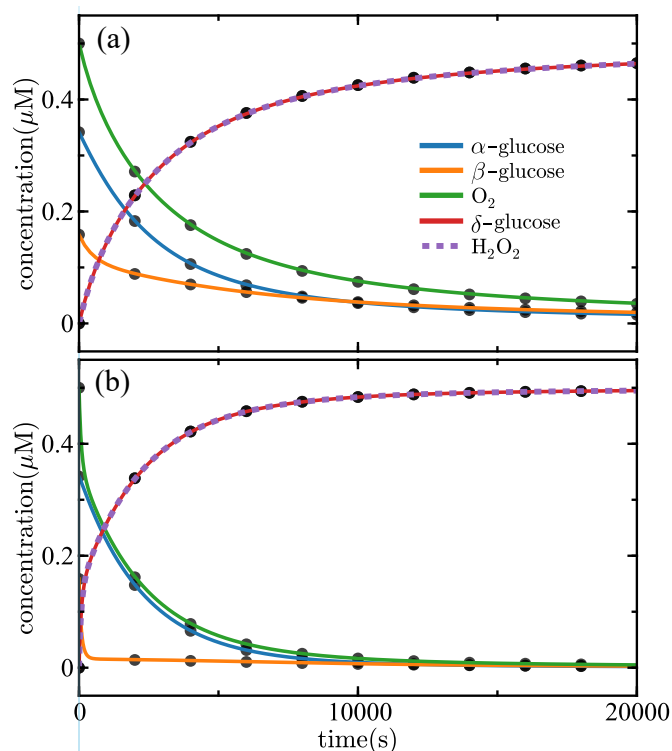
With respect to the performance of SISR on fast/slow systems, three points are of note: (1) Compared to coefficient thresholding-based methods in the literature (such as SINDy-based approaches), SISR appears better able to extract fast/slow dynamics, (2) samples are taken in the fast region in the training data which is why the SISR method can recover the dynamics in that region, however inference in the absence of training data may capture these dynamics, and (3) further investigation on the use of SISR for fast/slow and stiff chemical kinetic systems is an important line of inquiry.

## 4 Conclusions

A stoichiometrically-informed symbolic regression (SISR) method was introduced to automate the discovery of chemical reaction mechanisms from time series chemical concentration data. The SISR method generates symbolic forms for reaction mechanisms, and will be particularly useful when examining complex chemical systems where manual derivation of a mechanism is difficult or simply impractical. By integrating a physics-informed mathematical framework that accounts for the intrinsic functional geometry of chemical reaction mechanisms and the stoichiometry of those mechanisms, the developed method successfully identified sparse and interpretable reaction mechanisms for multiple example chemical processes. Specifically, the SISR results demonstrated strong agreement between predicted and true mechanisms across a range of chemical reaction schemes.

Compared to existing approaches for symbolic dynamical system discovery such as SINDy and its variants<sup>40–42</sup>, the SISR





**Fig. 15** Time evolution of the concentrations of chemical species in the Glucose Oxidation reaction mechanism given in Eq. (48). Panel (a) shows the case with  $k_2 = 2.78 \times 10^{-3} \mu\text{M}^{-2}\text{s}^{-1}$  and panel (b) shows the case with  $k_2 = 2.78 \times 10^{-4} \mu\text{M}^{-2}\text{s}^{-1}$ . The solid lines are the results of the SISR method and the corresponding black markers are a subset of the data used by SISR to extract the reaction mechanism and fit the rate constants. The concentrations are shown in micromolar and time is shown in seconds.

method alleviates key limitations by (a) incorporating stoichiometric constraints, (b) reducing the reliance on predefined reaction assumptions, and (c) improving robustness in handling fast-slow dynamics due to the reduction of hyperparameter tuning and thresholding procedures. While SINDy-based approaches can offer significant computational advantages in comparison to SISR, for situations in which physical insight, interpretability, accuracy, and the ability to forecast future system states are important, SISR offers strong advantages. All the SISR computational operations presented in this work can be performed on personal computers with standard commercial hardware. Generally and broadly, we found that to complete the SISR search over the defined spaces would take hours to a day in wall time. This computational speed will be sufficient for analysis of many datasets. When a large number of species are included in the data, the SISR process can be scaled up simply using parallel computing. For example, for each generation in the genetic algorithm, the evaluation of each candidate mechanism on each island is not correlated with the other candidate mechanisms, and therefore the search process can be parallelized.

A key limitation of the current approach is that the rate constants extracted from the data are purely numerical values and not symbolic functional forms that can be applied to account for changes in environmental and/or thermodynamic conditions such as changes in temperature or pressure. This can be a problem if the derived mechanism will be used to forecast the outcome of the chemical process at a thermodynamic state other than the state used to generate/measure the data. Additionally, due to this limitation, nonequilibrium cases in which the rate constants are varying in time as parametric coefficients<sup>12,48,75</sup> or over thermodynamic conditions are currently outside the scope of the current approach. Future work will address nonequilibrium chemical reaction mechanisms<sup>14,76,77</sup>. Improving the computational efficiency of the method using different optimization strategies is an important next step. Additionally, validation across a broad set of experimental datasets would further establish the utility of SISR on noisy, incomplete, and complex data, and work in this direction is currently underway.

## Conflicts of interest

There are no conflicts to declare.

## Data availability

The code and data that support the findings of this study are available at [www.github.com/rxhernandez/SISR](https://github.com/rxhernandez/SISR) and <https://doi.org/10.5281/zenodo.19577107>.

## Acknowledgments

We acknowledge support from the Los Alamos National Laboratory (LANL) Directed Research and Development funds (LDRD). We also acknowledge support to RH and to MPB from the National Science Foundation grant No. 2102455. The computing resources used to perform this research were provided in part by the LANL Institutional Computing Program, and by the Advanced Research Computing at Hopkins (ARCH) high-performance computing (HPC) facilities.



## Notes and references

- M. R. Harper, K. M. Van Geem, S. P. Pyl, G. B. Marin and W. H. Green, *Combustion and Flame*, 2011, **158**, 16–41.
- H. Zhu, R. J. Kee, V. M. Janardhanan, O. Deutschmann and D. G. Goodwin, *J. Electrochem. Soc.*, 2005, **152**, A2427.
- H. Gossler, L. Maier, S. Angeli, S. Tischer and O. Deutschmann, *Catalysts*, 2019, **9**, 227.
- R. Abramovitch, E. Tavor, J. Jacob-Hirsch, E. Zeira, N. Amariglio, O. Pappo, G. Rechavi, E. Galun and A. Honigman, *Cancer Research*, 2004, **64**, 1338–1346.
- S. Stocker, G. Csanyi, K. Reuter and J. T. Margraf, *Nature Communications*, 2020, **11**, 5505.
- J. Burés and I. Larrosa, *Nature*, 2023, **613**, 689–695.
- Q. Zhao and B. M. Savoie, *Nature Computational Science*, 2021, **1**, 479–490.
- Q. Zhao and B. M. Savoie, *Proc. Natl. Acad. Sci. U.S.A.*, 2023, **120**, e2305884120.
- Y. Mizuno and T. Komatsuzaki, *Phys. Rev. Res.*, 2024, **6**, 013115.
- Y. Matsumura, K. Tabata and T. Komatsuzaki, *J. Chem. Theory Comput.*, 2025, **21**, 3523–3535.
- C.-S. Liu, R. Hernandez and G. B. Schuster, *J. Am. Chem. Soc.*, 2004, **126**, 2877–2884.
- D. Yang, K. Chen, A. S. Edgar, I. Matanovic, J. Jung and J. D. Kress, *Polym. Degrad. Stab.*, 2023, **218**, 110565.
- E. Pollak and P. Talkner, *Chaos*, 2005, **15**, 026116.
- M. Z. Bazant, *Acc. Chem. Res.*, 2013, **46**, 1144–1160.
- G. T. Craven and R. Hernandez, *Phys. Rev. Lett.*, 2015, **115**, 148301.
- G. T. Craven and A. Nitzan, *Proc. Natl. Acad. Sci.*, 2016, **113**, 9421–9429.
- D. V. Matyushov, *Proc. Natl. Acad. Sci.*, 2016, **113**, 9401–9403.
- D. M. Wilary and J. M. Cole, *Journal of Chemical Information and Modeling*, 2023, **63**, 6053–6067.
- M. de Carvalho Servia, I. O. Sandoval, K. K. M. Hii, K. Hellgardt, D. Zhang and E. Antonio del Rio Chanona, *Digital Discovery*, 2024, **3**, 954–968.
- J. Huang, Y. Zhou and W.-A. Yong, *Journal of Computational Physics*, 2022, **448**, 110743.
- G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto and L. Zdeborová, *Rev. Mod. Phys.*, 2019, **91**, 045002.
- S. Zhong, K. Zhang, M. Bagheri, J. G. Burken, A. Gu, B. Li, X. Ma, B. L. Marrone, Z. J. Ren, J. Schrier *et al.*, *Environ. Sci. and Technol.*, 2021, **55**, 12741–12754.
- A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero and S. Drăghici, *PLoS Comp. Biol.*, 2007, **3**, e116.
- J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. De Fabritiis, F. Noé and C. Clementi, *ACS Cent. Sci.*, 2019, **5**, 755–767.
- M. Welborn, L. Cheng and T. F. Miller III, *J. Chem. Theory Comput.*, 2018, **14**, 4772–4779.
- M. Kulichenko, J. S. Smith, B. Nebgen, Y. W. Li, N. Fedik, A. I. Boldyrev, N. Lubbers, K. Barros and S. Tretiak, *J. Phys. Chem. Lett.*, 2021, **12**, 6227–6243.
- K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547.
- J. Carrasquilla and R. G. Melko, *Nature Phys.*, 2017, **13**, 431.
- G. Carleo and M. Troyer, *Science*, 2017, **355**, 602–606.
- L. Johnson, W. Malone, J. Rizk, R. Chen, T. Gibson, M. W. Cooper and G. T. Craven, *Compl. Matls. Sci.*, 2024, **242**, 113079.
- J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe and S. Lloyd, *Nature*, 2017, **549**, 195.
- D.-L. Deng, X. Li and S. Das Sarma, *Phys. Rev. X*, 2017, **7**, 021021.
- Y. Liu, W. Hong and B. Cao, *Energy*, 2019, **188**, 116091.
- G. T. Craven, N. Lubbers, K. Barros and S. Tretiak, *J. Phys. Chem. Lett.*, 2020, **11**, 43724378.
- G. T. Craven, N. Lubbers, K. Barros and S. Tretiak, *J. Chem. Phys.*, 2020, **153**, 104502.
- R. Jiang, P. Singh, F. Wrede, A. Hellander and L. Petzold, *PLoS computational biology*, 2022, **18**, e1009830.
- S.-M. Udrescu and M. Tegmark, *Sci. Adv.*, 2020, **6**, eaay2631.
- D. Angelis, F. Sofos and T. E. Karakasidis, *Arch. Comput. Methods Eng.*, 2023, **30**, 3845–3865.
- Y. Wang, N. Wagner and J. M. Rondinelli, *MRS Communications*, 2019, **9**, 793805.
- S. L. Brunton, J. L. Proctor and J. N. Kutz, *Proc. Natl. Acad. Sci.*, 2016, **113**, 3932–3937.
- M. Hoffmann, C. Fröhner and F. Noé, *J. Chem. Phys.*, 2019, **150**, 025101.
- N. Bhatt, B. Jayawardhana and S. S.-E. Plaza, 2023 62nd IEEE Conference on Decision and Control (CDC), 2023, pp. 3512–3518.
- K. Kaheman, J. N. Kutz and S. L. Brunton, *Proc. R. Soc. A*, 2020, **476**, 20200279.
- B. M. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J. N. Kutz and S. L. Brunton, *J. Open Source Softw.*, 2020, **5**, 2104.
- B. Prokop and L. Gelens, *Isience*, 2024, **27**, 109316.
- A. Sandoz, V. Ducret, G. A. Gottwald, G. Vilmart and K. Peron, *Proc. R. Soc. A*, 2023, **479**, 20220556.
- B. Prokop, N. Frolov and L. Gelens, *Chaos*, 2024, **34**, 063135.
- S. Li, E. Kaiser, S. Laima, H. Li, S. L. Brunton and J. N. Kutz, *Phys. Rev. E*, 2019, **100**, 022220.
- M. Dam, M. Brøns, J. Juul Rasmussen, V. Naulin and J. S. Hesthaven, *Phys. Plasmas*, 2017, **24**, 022310.
- E. Sakamoto and H. Iba, Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546), 2001, pp. 720–726.
- M. Quade, M. Abel, K. Shafi, R. K. Niven and B. R. Noack, *Phys. Rev. E*, 2016, **94**, 012214.
- R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Mater.*, 2018, **2**, 083802.
- T. A. R. Purcell, M. Scheffler and L. M. Ghiringhelli, *J. Chem. Phys.*, 2023, **159**, 114110.



- 54 J. J. Bramburger, D. Dylewsky and J. N. Kutz, *Phys. Rev. E*, 2020, **102**, 022204.
- 55 G. Ducci, M. Kouyate, K. Reuter and C. Scheurer, *The Journal of Chemical Physics*, 2025, **162**, 114118.
- 56 Z. T. Wilson and N. V. Sahinidis, *Computers & Chemical Engineering*, 2019, **127**, 88–98.
- 57 W. Chen, L. T. Biegler and S. García-Muñoz, *AIChE Journal*, 2018, **64**, 3595–3613.
- 58 W. Bradley, G. S. Gusmão, A. J. Medford and F. Boukouvala, *14th International Symposium on Process Systems Engineering*, Elsevier, 2022, vol. 49, pp. 1741–1746.
- 59 G. S. Gusmão, A. P. Retnanto, S. C. da Cunha and A. J. Medford, *Catalysis Today*, 2023, **417**, 113701.
- 60 S. Prabhu, N. Kosir, M. V. Kothare and S. Rangarajan, *Industrial & Engineering Chemistry Research*, 2025, **64**, 2601–2615.
- 61 Q. Wu, T. Avanesian, X. Qu and H. V. Dam, *J. Chem. Phys.*, 2022, **157**, 164801.
- 62 M. R. Muthyala, F. Sorourifar, Y. Peng and J. A. Paulson, *Industrial & Engineering Chemistry Research*, 2025, **64**, 3354–3369.
- 63 J. Lee, L. J. Augustine, G. Henkelman, P. Yang and D. Perez, *J. Chem. Theory Comput.*, 2025, **21**, 5182–5193.
- 64 B. M. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J. N. Kutz and S. L. Brunton, *arXiv preprint arXiv:2004.08424*, 2020.
- 65 The scaling factor is not applied in cases where the maximum concentration is zero.
- 66 M. Kommenda, A. Beham, M. Affenzeller and G. Kronberger, in *Complexity Measures for Multi-objective Symbolic Regression*, Springer International Publishing, 2015, p. 409416.
- 67 N. Haut, Z. Huang and A. Alessio, *arXiv preprint arXiv:2501.17372*, 2025.
- 68 G. F. Smits and M. Kotanchek, *Genetic programming theory and practice II*, 2005, 283–299.
- 69 W. Bechtel and A. Abrahamsen, *Philosophy of complex systems*, Elsevier, 2011, pp. 257–285.
- 70 D. Soloveichik, G. Seelig and E. Winfree, *Proceedings of the National Academy of Sciences*, 2010, **107**, 5393–5398.
- 71 P. Zheng, T. Askham, S. L. Brunton, J. N. Kutz and A. Y. Aravkin, *IEEE Access*, 2019, **7**, 1404–1423.
- 72 K. A. Johnson and R. S. Goody, *Biochemistry*, 2011, **50**, 8264–8269.
- 73 A. Cornish-Bowden, *Perspectives in Science*, 2015, **4**, 3–9.
- 74 Z. Tao, R. A. Raffel, A.-K. Soud and J. Goodisman, *Biophysical journal*, 2009, **96**, 2977–2988.
- 75 S. Rudy, A. Alla, S. L. Brunton and J. N. Kutz, *SIAM Journal on Applied Dynamical Systems*, 2019, **18**, 643–660.
- 76 G. T. Craven and R. Hernandez, *Phys. Chem. Chem. Phys.*, 2016, **18**, 4008–4018.
- 77 Y. Nagahata, R. Hernandez and T. Komatsuzaki, *J. Chem. Phys.*, 2021, **155**, 210901.



The code and data that support the findings of this study are available at [www.github.com/rxhernandez/SISR](https://www.github.com/rxhernandez/SISR) and <https://doi.org/10.5281/zenodo.19577107>.

