Computers and Chemical Engineering 202 (2025) 109265

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

DoE-SINDy: an automated framework for model generation and selection in kinetic studies

Wenyao Lyu, Federico Galvanin [*]

Department of Chemical Engineering, University College London (UCL) Torrington Place, London WC1E 7JE, United Kingdom

ARTICLE INFO

Keywords:
Model identification
Model generation
Kinetic studies
Sparse regression
Data-driven modelling

ABSTRACT

Efficient and accurate identification of kinetic models is critical for understanding chemical reaction mechanisms and enabling process optimisation and control. This study introduces DoE-SINDy, an enhanced framework that integrates design of experiments (DoE) with the Sparse Identification of Nonlinear Dynamics (SINDy) methodology to improve the reliability and interpretability of identified models under constraints of noisy, sparse and small experimental datasets. Unlike existing approaches, DoE-SINDy employs experimental-level subsampling for model generation, which reduces the inclusion of biased trajectories and ensures the identified model is representative. The framework further incorporates parameter re-estimation, non-significant terms removal, and identifiability analysis to enhance model robustness, reduce complexity, and reject overly complex or unidentifiable models. Rigorous model evaluation and selection steps, guided by flexible stopping criteria, strike a balance between statistical accuracy and computational efficiency. The methodology is tested on a simulated batch-reaction case study. DoE-SINDy consistently outperforms original SINDy and ensemble-SINDy (ESINDy) in recovering ground-truth models and achieving convergence to optimal structures as the dataset grows.

1. Introduction

Digitalisation is revolutionising chemical engineering by enabling advanced technologies like digital twins, which provide real-time virtual representations of physical systems. These tools enhance monitoring, optimisation, and decision-making (Javaid et al., 2023), relying heavily on robust kinetic models to describe the behaviour of chemical and biochemical reaction systems under various operating conditions. Accurate models support improved control schemes, product quality, and production rates (Mclean and Mcauley, 2012). However, identifying such models remains challenging as it requires extensive and time-consuming experimentation and costly numerical analytical resources.

A key challenge in developing digital twins is the identification of the model structure, including determining the appropriate set of equations, and precise estimation of the model-specific parameters. This task is particularly complex for chemical and biochemical reaction systems. Numerous intermediates and by-products are very difficult to describe kinetically in detail. Additionally, limited variable observability, incomplete data and unavoidable experimental noise further exacerbate the challenge (Maria, 2004). These difficulties highlight the need for a systematic framework capable of efficiently identifying kinetic models using limited and noisy data while minimising experimental demands.

The processes investigated in this work are assumed to be represented by mathematical expressions and associated parameters that well determine the relationship between state variables within the physical system and represent the dynamic behaviour of a process through deterministic models (Bard, 1974) typically formulated through a set of differential and algebraic equations (DAEs).

The reliability of a kinetic model is represented by both model accuracy and parameter precision. Model accuracy measures how well a model predicts system behaviour based on the data, while parameter precision reflects the uncertainty of the parameter estimates. An accurate model requires a model structure that accurately represent the system's underlying kinetics and minimum parameter uncertainty.

Model structure confirmation must precede parameter estimation and validation, as an incorrect structure introduces bias and undermines reliability. Existing model structure identification approaches can be broadly classified into three main categories: 1) model selection; 2) model modification and 3) model generation.

In model selection approaches the most suitable model structure is identified from a set of candidate models using statistical metrics

[*] Corresponding author.
E-mail address: f.galvanin@ucl.ac.uk (F. Galvanin).

https://doi.org/10.1016/j.compchemeng.2025.109265
Received 31 January 2025; Received in revised form 21 June 2025; Accepted 24 June 2025
Available online 28 June 2025

obtained after data fitting. These candidate models may be derived from first principles or sourced from existing literature. Model selection is embedded in model-building approaches, as proposed by Asprey and Macchietto (2000), by employing a sequential process involving identifiability analysis (Vanrolleghem et al., 1995; Miao et al., 2011; Dobre et al., 2012; Deussen and Galvanin, 2022; Binns et al., 2024; Sangoi et al., 2025), model-based design of experiments (MBDoE) for model discrimination (Bawa et al., 2023; Buzzi-Ferraris and Forzatti, 1983; Hunter and Reiner, 1965; Schaber et al., 2014; Schwaab et al., 2006; Sen et al., 2021), and MBDoE for improving parameter precision (Asprey et al., 2000; Franceschini and Macchietto, 2008; Galvanin et al., 2007; Gottu Mukkula and Paulen, 2019; López C. et al., 2015; Quaglio et al., 2019; Schaber et al., 2014) to systematically select the most suitable model structure and ensure precise parameter estimation. Superstructure-based approaches (Edwards et al., 2000; Tsay et al., 2017) formulate a superstructure comprising all potential component functions, each associated with a binary decision variable to selectively activate corresponding mode, thereby addressing the model-identification problem as a mixed-integer optimisation. Superstructure-based approaches allow reliable parameter estimation without requiring pre-selection of the optimal model structure. Artificial neural network (ANN)-based approaches bypass the need for extensive parameter fitting by training classifiers with in-silico data from candidate models to directly identify optimal structures, as demonstrated by Quaglio et al. (2020b), and can be enhanced through MBDoE to improve classification accuracy and experimental efficiency (Sangoi et al., 2024).

When all candidates do not match the system observations or are poorly identifiable, **model modification approaches** can guide the refinement of existing structures or parameterisation according to specific metrics. Ill-conditioning in system identification often arises when the model structure poorly matches the system or when the available data lacks sufficient informativeness. Regularisation methods are commonly used for addressing such challenges by introducing constraints or penalties to attract the excessive degrees of freedom towards reasonable values (Tikhonov and Arsenin, 1977), such as ridge regression, principal component regression (PCR), and Tikhonov regularisation (Johansen, 1997; Sjöberg et al., 1993; Sjöberg and Ljung, 1992). Incremental modelling iteratively adjusts model complexity through diagnostic-driven refinement and experimental validation. Parameter subset selection (SsS) methods (Barz et al., 2013) integrated with experimental design, can be employed to incrementally fix poorly identifiable parameters. Meneghetti et al. (2014) proposed a methodology to diagnose model equations or parameters responsible for the mismatch by comparing the correlation structures of historical and simulated datasets using principal component analysis (PCA), without requiring additional experiments. Quaglio et al. (2020a) developed an iterative model-building framework based on maximum likelihood inference to achieve an appropriate level of complexity indicated by the introduced model modification indexes (MMIs). Hybrid modelling, which combines first-principles equations with data-derived components, also addresses mismatches effectively (Molga and Cherbański, 1999; Zhang et al., 2020; Narayanan et al., 2022; Schweidtmann et al., 2024; Jul-Rasmussen et al., 2024; Chen and Ierapetritou, 2020).

Model generation approaches focus on identifying model structures from observed measurements, providing a foundation for subsequent parameter estimation, particularly in cases where prior structural knowledge is limited. These methods emphasise model explicability and interpretability, which are vital for building trust in physical and engineering sciences (Venkatasubramanian, 2019). Evolutionary algorithms, including genetic programming (GP), optimises model structures and parameters by mimicking natural selection, ensuring physical interpretability by incorporating domain-specific constraints and priors, such as elementary functions and transformations rooted in underlying physiochemical mechanisms (Chakraborty et al., 2021). Symbolic regression (SR), rooted in GP, builds interpretable models by optimizing mathematical expression trees that combine operands and operators (Angelis et al., 2023). Sparse regression techniques focus on deriving concise, interpretable models that balance accuracy and complexity. Sparse identification of nonlinear dynamics (SINDy) is a one of the main sparse-regression approaches that is capable of deriving ODE/PDE models with minimal prior knowledge of physical mechanisms (Brunton et al., 2016). Sparse Shooting S is a fast, cellwise robust regression method that excels in prediction problems with more predictor variables than observations and outliers (Bottmer et al., 2022).

SINDy and its numerous extensions have become prominent for sparse regression in scientific discovery, enhancing its feasibility across diverse scenarios. Generalisations like implicit-SINDy and SINDy-PI extend its applicability to systems with rational function nonlinearities (Mangan et al., 2016; Kaheman et al., 2020), stochastic SINDy addresses stochastic systems (Boninsegna et al., 2018), and OASIS framework extends SINDy for adaptive, automatic and accurate modelling during rapid dynamics changes (Bhadriraju et al., 2020). Enhancements such as SR3 improve computational efficiency and flexibility (Zheng et al., 2019). Techniques like PINN-SR (Chen et al., 2021), WSINDy (Messenger and Bortz, 2021), Dropout-SINDy (Abdullah et al., 2022a), and SISC (Abdullah et al., 2022b) enhance noise robustness and scalability. Trapping SINDy constrains models within boundaries by integrating a global stability theorem (Kaptanoglu et al., 2021). Ensemble-SINDy merge multiple models to improve prediction accuracy and leverage active learning to reduce data demands (Fasel et al., 2022). Modified SINDy (Kaheman et al., 2022) and EKF-SINDy (Rosafalco et al., 2024) enhance the noise robustness and computational efficiency by integrating automatic differentiation and noise quantification.

Despite their advancements, the three main model identification approaches face notable limitations, particularly in complex chemical and biochemical processes. Model selection and modification methods heavily rely on prior knowledge, such as explicit kinetic models or mechanistic structures, which are often unavailable. Model generation approaches, while more flexible by leveraging libraries of candidate functions, encounter challenges including high variability under different experimental conditions, lack of rigorous evaluation and selection stages, and potential generation of unidentifiable models or models lacking physical or mechanistic meaning.

Focussing on the limitations of the lack of identifiability analysis and insufficient model evaluation and aiming at increasing the success rate of the identification of ‘true’ model from small, sparse and noisy data, we propose the Design of Experiments Integrated Sparse Identification of Nonlinear Dynamics (DoE-SINDy) framework. This framework applied subsampling technique during model generation to increase the likelihood of identifying the ‘true’ model structure. By incorporating identifiability analysis, model calibration, validation and selection steps, DoE-SINDy ensures the reliability, robustness, and accuracy of the identified models, even from limited and noisy datasets.

The article is structured as follows. Section 2 introduces the methodology of DoE-SINDy, detailing its framework, including design of experiments, model generation, ranking, calibration, simplification, and validation steps, along with the criteria for assessing both the identified models and the overall performance of model identification approaches. Section 3 presents a case study, describing the generation of in-silico data and the implementation of DoE-SINDy for model identification. Section 4 discusses the results, focusing on a performance comparison of the models identified by SIDNy, ESINDy, and DoE-SINDy, as well as their success rates. Finally, Section 5 concludes the article by summarising key findings and proposing directions for future work.

2. Methodology

We assume that the dynamic behaviour of a process system is mathematically represented by a deterministic model (Bard, 1974; Quaglio et al., 2020a), typically formulated through a set of differential and algebraic equations (DAEs). These equations can be expressed generally as:

$$\left\{ \mathbf{f}(\dot{\mathbf{x}},\mathbf{x},\mathbf{u},t,\boldsymbol{\theta}) = 0\hat{\mathbf{x}} = \mathbf{h}(\mathbf{u},\ t,\ \boldsymbol{\theta}) \right. \tag{1}$$

Here, $\mathbf{f}$ and $\mathbf{h}$ are respectively an N_f and N_y – dimensional vector of model equations, $\mathbf{x}$ is an N_x – dimensional vector of state variables, $\mathbf{u} \in$ U is an N_u – dimensional vector of control input variables, t is time and $\boldsymbol{\theta}$ is an N_θ– dimensional vector of model parameters. The function $\mathbf{h}$ serves as a selection operator that determines which variables are measured from the system, while $\hat{\mathbf{x}}$ represents an N_x – dimensional vector of predictions for the measurable set of system state variables. In the context of model identification, the objective is to determine the function $\mathbf{f}$ that describes the relationship between the state variables $\mathbf{x}$, their time derivatives $\dot{\mathbf{x}}$, system inputs $\mathbf{u}$ and other system quantities over time (t). This involves identifying both the model structure (the functional form of each equation) and the corresponding model-specific parameter set $\boldsymbol{\theta}$.

In this research, we aim to develop a systematic framework for simultaneously identifying both the model structure and parameters, ensuring reliability and interpretability. To achieve this, we propose an iterative model identification framework named Design of Experiments Integrated Sparse Identification of Nonlinear Dynamics (DoE-SINDy).

2.1. DoE-SINDy framework

DoE-SINDy systematically generates and selects the most suitable model for representing dynamic processes, which address variability of model generation in original SINDy, ensuring the identifiability and enhancing the reliability of the identified models. The structure and workflow of the proposed framework are illustrated in Fig. 1.

A key limitation of the original SINDy algorithm is that it discovers a model only once, working with a fixed dataset and assuming that the resulting structure will remain valid under any additional operating conditions. In practice, dynamic-process data collected under different experimental settings, such as varying initial concentrations in a reaction network, often produce markedly different trajectories. Therefore, the structure identified by SINDy can be highly sensitive to the specific data used for training and may fail to generalise when new experiments are performed. New data may even introduce or remove terms, leading to model instability and reduced predictive power. DoE-SINDy addresses this issue by adopting an iterative identification strategy: with each cycle the framework adds newly generated experiments to the training pool, regenerates candidate structures, and recalibrates parameters. This progressive refinement promotes structural consistency and robustness across a wide range of operating conditions, ultimately yielding models with improved generalisation and interpretability.

2.1.1. Step 1 and 2: design of experiments and data collection

Identifying dynamic models via DoE-SINDy requires multiple trajectories of time-series state variables. Thus, design of experiment (DoE) techniques, such as Latin hypercube sampling (LHS) (McKay et al., 1979) and uniform sampling (Virtanen et al., 2020), are incorporated to design multiple sets of operating conditions providing as much information as possible while constrained within physical limits. The common design factors in chemical and biochemical processes are temperature, pressure, initial concentration (batch systems) or feed specification (flow systems), residence time and sample size per experiment.

To ensure the minimum cost and time in experimentation, DoE-SINDy is initialised with a small dataset and incrementally expanded

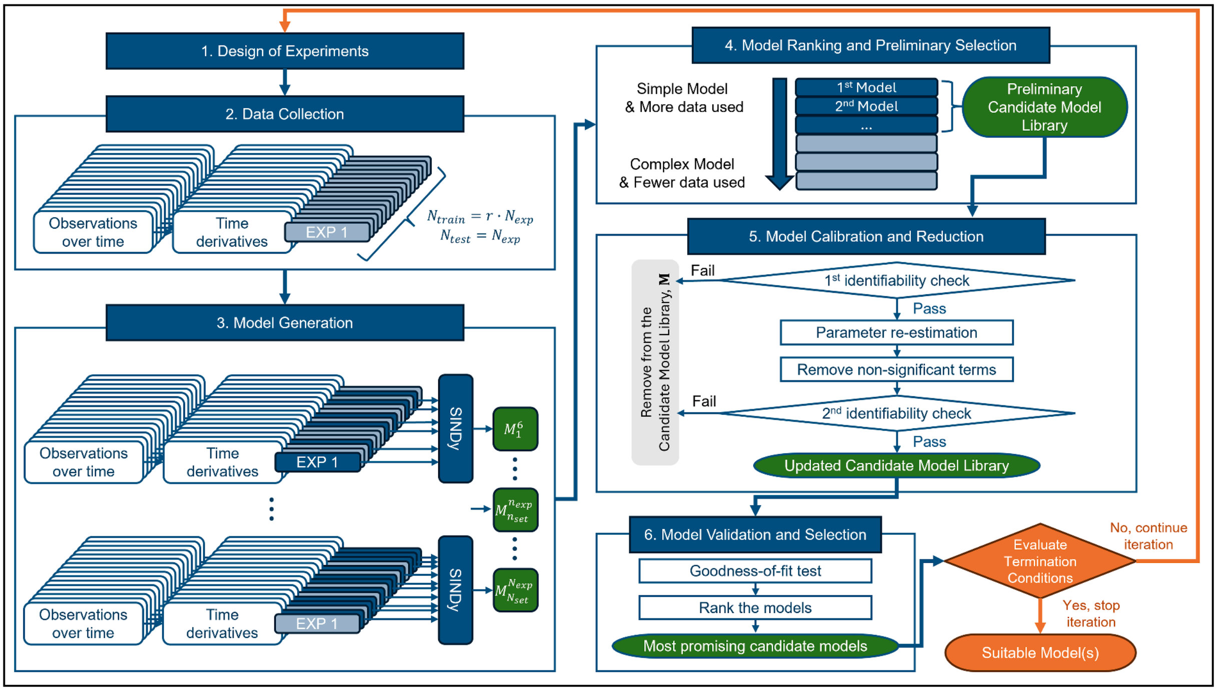

1. Design of Experiments
2. Data Collection
Observations over time
Time derivatives
EXP 1
$N_{train} = r \cdot N_{exp}$
$N_{test} = N_{exp}$
3. Model Generation
Observations over time
Time derivatives
EXP 1
SINDy
M_1^6
M_{Nset}^{nexp}
Observations over time
Time derivatives
EXP 1
SINDy
M_{Nset}^{Nexp}
4. Model Ranking and Preliminary Selection
Simple Model & More data used
Complex Model & Fewer data used
1st Model
2nd Model
...
Preliminary Candidate Model Library
5. Model Calibration and Reduction
Remove from the Candidate Model Library, M
Fail
1st identifiability check
Pass
Parameter re-estimation
Remove non-significant terms
Fail
2nd identifiability check
Pass
Updated Candidate Model Library
6. Model Validation and Selection
Goodness-of-fit test
Rank the models
Most promising candidate models
Evaluate Termination Conditions
No, continue iteration
Yes, stop iteration
Suitable Model(s)

Fig. 1. DoE-SINDy framework for identifying the most suitable model(s) from experimental data. The DoE-SINDy framework begins with a preliminary design of experiments (DoE) to explore experimental conditions constrained by physical limits (step 1 in Fig. 1). In the second step, measurements of state variables are collected and used to numerically approximate their time derivatives, which are then split into training and validation subsets (step 2). Multiple candidate models are generated from subsets of the training dataset using original SINDy in step 3. In step 4, these candidates are then ranked first by complexity, measured by the number of non-zero coefficients, and within each complexity group, by the number of experiments used, prioritising simpler, well-supported models for preliminary selection. Model calibration follows, in step 5, incorporating parameter re-estimation using the full training set and refinement to remove non-significant terms. Identifiability analysis is conducted before and after calibration to retain only identifiable models. Validated models are then evaluated against the user-defined stopping criterion, either ‘and’, ‘chi2’, ‘normality’ or ‘or’ in step 6. If no model meets the criteria, the framework iteratively updates the experimental design and expands the dataset. Iteration continues until a model satisfies the criteria or the experimental budget is depleted. The final output consists of the most statistically acceptable model(s), ranked by the Akaike Information Criterion (AIC).

until user-defined stopping criteria (defined in Section 2.1.5) or budget limitations are reached. Before starting DoE-SINDy, a fixed pool of designs is generated in advance using either LHS or uniform sampling, based on the predefined budget. In each iteration, one additional design is randomly selected from the remaining unused designs in this candidate pool to collect additional data.

Thus, the starting number of experiments and the upper limit of the number of experiments is another important argument for design of experiments. In DoE, the minimum number of experiments is often chosen based on the model to be calibrated. For instance, when fitting a linear model with interactions, a common starting point is to set the minimum number of experiments to twice the number of factors (Hone et al., 2019). The maximum number of experiments to be conducted can depend on the experimental budget.

The state measurements $x_1, \ldots, x_{N_v}$ collected from the experiments $e_1, \ldots, e_{N_{\text{exp}}}$ sampled at time step $t_1^{e_i}, \ldots, t_{N_m}^{e_i}$ are organised into the data matrix $\mathbf{x}$ with dimensions $(N_{\text{exp}} \cdot N_m) \times N_v$,

$$\mathbf{x} = \begin{bmatrix} x_1(t_1^{e_1}) & \cdots & x_n(t_1^{e_1}) \\ \vdots & \ddots & \vdots \\ x_1\left(t_{N_m}^{e_{N_{\text{exp}}}}\right) & \cdots & x_n\left(t_{N_m}^{e_{N_{\text{exp}}}}\right) \end{bmatrix} \tag{2}$$

The derivatives of the state variables along these trajectories are numerically approximated. The choice of numerical differentiation method plays a critical role in the accuracy and reliability of the identified model, particularly when dealing with sparse or noisy data. The Python package ***derivative*** is employed to compute numerical derivatives from time-series data. This package provides a robust suite of differentiation methods tailored for various noise levels and data characteristics, including symmetric finite difference, Savitzky-Golay derivatives, spectral derivatives, spline derivatives of arbitrary order, total variation derivatives, Kalman derivatives, and kernel-based derivatives (Ahnert and Abel, 2007; Chartrand, 2011; Kaptanoglu et al., 2022; Silva et al., 2020; Tibshirani and Taylor, 2011). The resulting numerical derivatives are organised into a derivative matrix $\dot{\mathbf{x}}$, which has the same dimensions as the data matrix $\mathbf{x}$:

$$\dot{\mathbf{x}} = \left[\dot{x}_1(t_1^{e_1})\cdots\dot{x}_n(t_1^{e_1}) \vdots \ddots \vdots \dot{x}_1\left(t_{N_m}^{e_{N_{\text{exp}}}}\right)\cdots\dot{x}_n\left(t_{N_m}^{e_{N_{\text{exp}}}}\right) \right] \tag{3}$$

Each element in matrix $\dot{\mathbf{x}}$ represents the derivative of a corresponding state variable from the data matrix $\mathbf{x}$, collected at a specific experimental time point.

Then, the collected data, including both the measurements and the numerically approximated derivatives, are divided into training and validation sets. The training dataset is used to generate candidate models and calibrate their parameters. A key feature of our framework is that model validation is performed on the **entire accumulated dataset**, not just on a reserved validation subset. In each iteration, the newly acquired experimental data are mandatorily included in the training set, while a predefined ratio of experiments is randomly picked from the existing dataset to form the rest of the training set. The remaining data are used only for validation. After model generation and calibration using the training dataset, the candidate models are evaluated on the full set of accumulated data. This allows for explicit assessment of the model's **interpolation** capabilities (on the training dataset) and **extrapolation** capabilities (on previously unseen data), ensuring that the identified model is both accurate and generalisable across a wide range of operating conditions.

2.1.2. Step 3: model generation

SINDy derives a set of ODEs $\mathbf{f}(\mathbf{x})$ by identifying a combination of terms from a user-defined candidate term library $\boldsymbol{g}(\mathbf{x}) = \left[g_1(\mathbf{x}),\ g_2(\mathbf{x}),\ \ldots,\ g_\gamma(\mathbf{x})\right]$ associated with coefficients $\boldsymbol{\Xi} = \left[\boldsymbol{\xi}_1, \boldsymbol{\xi}_2,\ \ldots,\ \boldsymbol{\xi}_\gamma\right]$. This combination enforces the relationship between matrices $\mathbf{x}$ and $\dot{\mathbf{x}}$, formulated as Eq. (4).

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \approx \sum_{j=1}^{\gamma} g_j(\mathbf{x}) \boldsymbol{\xi}_j \tag{4}$$

To ensure both accuracy and extrapolation capability, the model discovery task is reformulated as an optimisation problem aiming at minimising the number of nonzero elements in the sparse coefficient matrix $\boldsymbol{\Xi}$. The package developed for SINDy algorithm, ***PySINDy v1.7.5***, is called for identifying single model in model generation section of DoE-SINDy (Kaptanoglu et al., 2022; Silva et al., 2020).

Fig. 2 illustrates and compares the schematic of generating a single model using SINDy, ESINDy and DoE-SINDy with different data-handling strategies. A model is generated using the entire training dataset without any subsampling (resampled without replacement) or bootstrapping (resampled with replacement) in SINDy (Brunton et al., 2016), as shown in Fig. 2a. ESINDy (Fasel et al., 2022) bootstraps (subsampled with replacement) data from every time-series trajectory, generating multiple ensemble SINDy models from different bootstrapped datasets, as illustrated in Fig. 2b These ensemble models are then aggregated by bagging (taking the mean of the identified coefficients) or bragging (taking the median of the identified coefficients). However, rather than subsampling point-wise data from every trajectory, DoE-SINDy subsamples entire experimental trajectories from the set of all conducted experiments and generates ensemble models from different subsets of experiments.

This design of experimental-level data subsampling in DoE-SINDy is tailored for chemical and biochemical kinetic studies, where experimental equipment often imposes constraints on sampling frequency. Derivatives are numerically approximated from the time series points (Section 2.1.1), so any change in the sampling grid directly affects their accuracy. For the subsampling method, more uncertainty is added due to the widened time step. The bootstrapping method estimates the missing values by interpolation, which is highly uncertain for sparse, nonlinear or rapidly changing trajectories. Thus, either point-wise subsampling or bootstrapping scheme therefore bias the subsequent sparse regression. DoE-SINDy therefore keeps every time point and instead retains complete trajectories and resamples at the level of whole experiments, preserving both temporal resolution and derivative fidelity.

Using design of experimental-level data subsampling instead of random point data subsampling prevents runs that unknowingly violate physical laws from distorting the model-identification process. We define an unknown physical constraint as any physical law or operating assumption that the regression step does not explicitly enforce. For example, temperature is high enough to activate a different reaction pathway. If a particular run violates such a rule, the data from that experiment no longer obeys the same kinetics and can distort structure identification. To guard against this, DoE-SINDy fits an ensemble of models, each built from a different random subset of complete experiments while preserving every time point within each retained run. The resulting candidates are subsequently calibrated and validated on the entire dataset. Any kinetic relationship that relies on a run violating an unseen constraint fails to appear consistently across the ensemble and is automatically discarded, leaving only the model that is both numerically sound and physically consistent.

As described in Section 2.1.1, the iterative approach in DoE-SINDy progressively incorporates additional experiments, up to the maximum number of experiments available. Within each iteration, subsets of experiments are generated by considering all possible combinations of experimental indices, starting from the minimum number defined for the initial iteration. This number increases incrementally up to the total number of experiments available. To ensure computational efficiency and diverse exploration of the experimental design space, either all

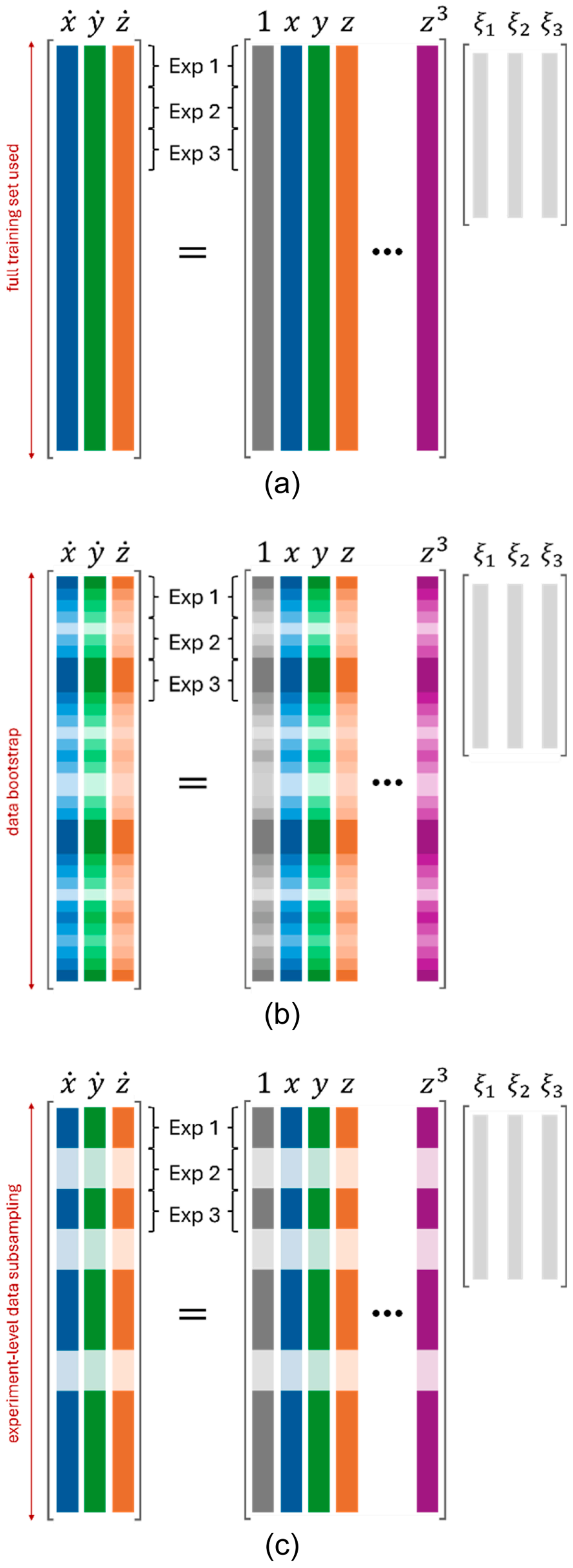


Fig. 2. Schematic of Eq. (4) in (a) SINDy (Brunton et al., 2016), (b) ESINDy (Fasel et al., 2022) and (c) DoE-SINDy. In (b) and (c), the colour bars indicate data usage: darker areas represent data used for model generation, while lighter areas represent unused data. In (b), the shading gradient reflects the density distribution of the bootstrapped samples over the points. In (c), the lighter segments correspond to entire experiments that are excluded from model generation via Eq. (4).

subsets or a specified number of subsets are selected randomly. These subsets are then fed into SINDy for model generation. Thus, during the model generation step, multiple models are created within a single iteration by leveraging all training dataset available up to that stage.

2.1.3. *Step 4: model ranking and preliminary selection*

In step 4 models generated in each iteration are ranked and preliminarily selected based on two key criteria: **simplicity** and **generalisation**. Simplicity is assessed by the model's complexity, quantified as the number of non-zero elements in the sparse coefficient matrix Ξ. Within each complexity group, models are further ranked by the number of experiments used in their derivation. This secondary criterion prioritises models supported by a broader range of experimental evidence, which is quantified by the number of experiments used for identification. Thus, it ensures their reliability on broader feasible operational regions and reduces the risk of overfitting.

In the preliminary selection step, models from the top-ranked complexity groups are selected, with a predefined number of models extracted from each group. By focusing on the most promising candidates—those that simultaneously exhibit simplicity and generalisation—this step conserves computational resources for the subsequent model calibration and reduction stages.

2.1.4. *Model calibration and reduction (Block 5 in Fig. 1)*

Regularisation methods such as LASSO and STLSQ are applied in model generation by shrinking small coefficients to zero, thereby controlling model complexity. They do not, however, test whether the retained parameters are uniquely identifiable. We therefore perform a separate identifiability analysis to ensure that the final model contains only parameters that can be precisely estimated.

Sensitivity-based practical identifiability analysis is implemented in DoE-SINDy framework before and after parameter re-estimation and non-significant terms removal to verify identifiability of the candidate models. The process begins by extracting the nonzero elements from the coefficient matrix Ξ to formulate a set of model parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_\eta]$. For example, if the approximated coefficient matrix is $\Xi = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 0 & 3 & 0 & 5 \end{bmatrix}$, the candidate parameter vector is $\boldsymbol{\theta} = [1, 2, 3, 5]$. Identifiability ensures that model parameters $\boldsymbol{\theta}$ can be uniquely determined from the system input $\mathbf{u}(t)$ and the observations $\mathbf{x}(t)$ (Miao et al., 2011). The sensitivity matrix $\mathbf{Q}$ quantifies how variations in parameters $\boldsymbol{\theta}$ influence the predicted trajectories of the model. Higher absolute values of its coefficients indicate greater parameter identifiability. In dynamic models, the sensitivity matrix varies with time, necessitating the use of a matrix of dynamic sensitivities $\mathbf{Q}$. Each element of $\mathbf{Q}$ corresponds to an instantaneous sensitivity matrix at a specific sampling point $t_m^{e_i}$ of an experiment e_i. The n, pth element of Q is calculated below as local first-order sensitivities of responses x_n to the parameters θ_p, typically approximated via the finite difference method:

$$\left[q\left(t_m^{e_i}\right)\right]_{np} = \frac{\left[\hat{x}\left(t_m^{e_i}\right)\right]_n(\theta_p + \varepsilon\theta_p) - \left[\hat{x}\left(t_m^{e_i}\right)\right]_n(\theta_p - \varepsilon\theta_p)}{2\varepsilon\theta_p} \tag{5}$$

where ε is a small perturbation imposed on parameters.

Fisher-information analysis is used here to diagnose local practical identifiability, i.e. whether at least one trajectory of the available data allows the unique estimation of every parameter (Waldron et al., 2019). For experiment e_i, the FIM profile $\mathbf{H}_\theta$ is calculated as:

$$[\mathbf{H}_\theta]_{kl} \cong \mathrm{H}_{\theta,initial} + \sum_{j=1}^{N_m} \frac{1}{\sigma_{ij}^2}\left[q\left(t_j^{e_i}\right)\right]_{nk}\left[q\left(t_j^{e_i}\right)\right]_{nl} \tag{6}$$

$$\mathrm{H}_{\theta,initial} = \frac{1}{12}(\theta_{ub} - \theta_{lb})^2 \mathrm{I} \tag{7}$$

If the $\mathbf{H}_\theta$ is full-rank, the model is practically identifiable; if $\mathbf{H}_\theta$ is singular or ill-conditioned, the model is practically unidentifiable with current data, even though it may be structurally identifiable in principle (Miao et al., 2011; Silvey, 1975). Only identifiable models are deemed suitable and structurally promising to subsequently target an improvement in parameter estimation.

In reaction systems with shared kinetic terms across multiple rate equations, such as those arising from reactions with identical stoichiometric coefficients, it is expected that these terms share the same kinetic parameter. However, sparse regression approaches, such as SINDy, lack mechanistic insights to link these terms, leading to the identification of independent coefficients to each term for the same physical contribution, and resulting in parameter redundancy. For example, a shared term representing a reaction rate appears with same value but opposite signs in the rate equations of a reactant and a product, reflecting stoichiometric relationships. If a model parameter is redundant, then it is not locally identifiable (Catchpole and Morgan, 1997) and the corresponding FIM matrix $\mathbf{H}_\theta$ detecting identifiability will be singular due to parameter redundancy.

Despite the limitation of parameter redundancy, such models should be retained because their structure is correct. Implementing reparameterisation techniques, such as parameter combinations, can solve this problem once the model structure is confirmed. To retain such models, pairwise comparison of the sensitivity profiles of any two coefficients is performed; if their profiles overlap completely (or are opposite) and do not overlap with the zero axis, the coefficients are recognised as paired and should be combined into a single parameter, as shown in Eq. (8).

$$s_k = \sum_{j=1}^{N_m} \left[q\left(t_j^{e_i}\right)\right]_{nk} \tag{8}$$

$$s_l = \sum_{j=1}^{N_m} \left[q\left(t_j^{e_i}\right)\right]_{nl}$$

$$\text{check if } s_k = s_l \ or\ s_k = -s_l \text{ and } s_k \neq 0,\ s_l \neq 0$$

In terms of the sensitivity profile $\mathbf{Q}$, only the row corresponding to one representative coefficient from each pair is retained. This adjustment ensures that the FIM matrix $\mathbf{H}_\theta$ does not become singular due to this issue, preventing such models being detected as unidentifiable and rejected.

Because both the original and modified versions of SINDy ignore measurement-noise statistics when fitting coefficients (Wei, 2022), the initial parameter values can be biased, and the selected structure can even be misidentified. To correct this bias, every model that passes the first identifiability check undergoes parameter re-estimation using the full training dataset via minimises the sum of weighted squared residuals, weighted by the inverse of the variance (Mandel, 1964):

$$\hat{\boldsymbol{\theta}} = \text{argmin} \sum_{n=1}^{N_v}\sum_{i=1}^{N_{\exp}}\sum_{j=1}^{N_m}\left(\frac{x\left(t_j^{e_i}\right) - x\left(t_j^{e_i}, \hat{\boldsymbol{\theta}}\right)}{\sigma}\right)^2 \tag{9}$$

Here, the standard deviation σ is assumed to be constant. This optimisation problem is solved using the Nelder-Mead method, implemented via the Python package *scipy.optimize.minimize* (Gao and Han, 2012; Virtanen et al., 2020). Thus, parameter re-estimation increases the accuracy of the model on a broader feasible region by calibrating parameters with more data and ensures robustness to the noise by weighting residuals to their variance. Fig. 3(a) and (b) compare the predictions of an example model against the observations of a trajectory before and after parameter re-estimation, showing significant improvements in fitting.

After parameter re-estimation, each calibrated model undergoes a non-significant-term removal step. The goal is to delete terms that contribute negligibly to the dynamics, thereby simplifying the structure without sacrificing predictive accuracy and ultimately improving reliability, interpretability, and generalisation. Each term in the model, defined as a candidate function multiplied by its estimated parameter, is evaluated for its contribution over time. The time-varying contributions of each term are computed using the predicted trajectories, and the total contribution of each term is aggregated across all time points. Terms with contributions below a predefined threshold ζ are deemed non-significant and removed by setting their corresponding parameters in the coefficient matrix Ξ to zero, as expressed in Eq. (10)

$$\text{if } \left| \sum_{i=1}^{N_{\exp}}\sum_{j=1}^{N_m} g_p\left(\mathbf{x},\ t_j^{e_i}\right)\hat{\xi}_{np} \right| \le \zeta, \ \text{ then } \hat{\xi}_{np} = 0 \tag{10}$$

As demonstrated in Fig. 3(b) and (c), the model maintains accurate fitting performance after removing terms with negligible contributions. This procedure ensures that the final model retains only essential components for accurately describing the system dynamics. Simplifying the model enhances interpretability and reduces computational complexity while capturing key system behaviours.

Finally, a second identifiability check is conducted on the refined models. The removal of non-significant terms may alter the model structure, necessitating re-evaluation to ensure that the parameters of the simplified model remain identifiable.

2.1.5. Step 6: model validation and selection

Calibrated models undergo validation to evaluate whether the identified models are statistically accurate enough to represent the system's dynamics.

A common goodness-of-fit test, two-tailed χ^2 test, detects the model accuracy by comparing the residual distribution against the hypothetical distribution of noise. From the assumption of the Gaussian distribution of the noise, the χ^2 test is conducted as Eq. (11) and (12) (Draper and Smith, 1998):

$$\chi^2 = 2\ \times \sum_{n=1}^{N_v}\sum_{i=1}^{N_{\exp}}\sum_{j=1}^{N_m}\left(\frac{x\left(t_j^{e_i}\right) - x\left(t_j^{e_i}, \hat{\boldsymbol{\theta}}\right)}{\sigma}\right)^2 \tag{11}$$

$$\begin{cases} \chi^2 < \chi^2_{ref}\left(\dfrac{1-\alpha}{2}\right) \ \text{Failed for overfitting} \\ \chi^2_{ref}\left(\dfrac{1-\alpha}{2}\right) < \chi^2 < \chi^2_{ref}\left(\dfrac{1+\alpha}{2}\right) \ \textit{Passed} \\ \chi^2 > \chi^2_{ref}\left(\dfrac{1+\alpha}{2}\right) \ \text{Failed for underfitting} \end{cases} \tag{12}$$

Failed for overfitting indicates that the model has an excessive number of parameters relative to the dataset, or the predictions are too close to the observations resulting in excessively low values for the χ^2 statistics. Underfitting indicates that the model structure fails to adequately describe the system. Conversely, a successful test confirms that the model adequately captures the dynamics while adhering to the Gaussian noise assumption.

For small dataset, passing χ^2 test could be challenging due to imprecise parameter estimates. As an alternative, a normality test is used to evaluate whether residuals follow a zero-mean Gaussian distribution (D'Agostino and Stephens, 1986). This less strict test allows the inclusion of models that may fail the χ^2 test due to parametric uncertainty.

If the goodness-of-fit test fails, the model is considered unsuitable. Four options for user-defined stopping criteria—passing both tests ('and'), passing only χ^2 test ('chi2'), passing only normality test ('normality'), or passing either ('or') —ranked from most to least strict, determine when the iteration of the DoE-SINDy procedure stops, with models passing the chosen criterion considered statistically acceptable. When all models in the current iteration fail to meet the stopping criterion, the framework generates new candidate models by expanding the dataset and iteratively refining the model library. This process continues until either a model satisfies the stopping criterion, or the experimental budget is fully utilised.

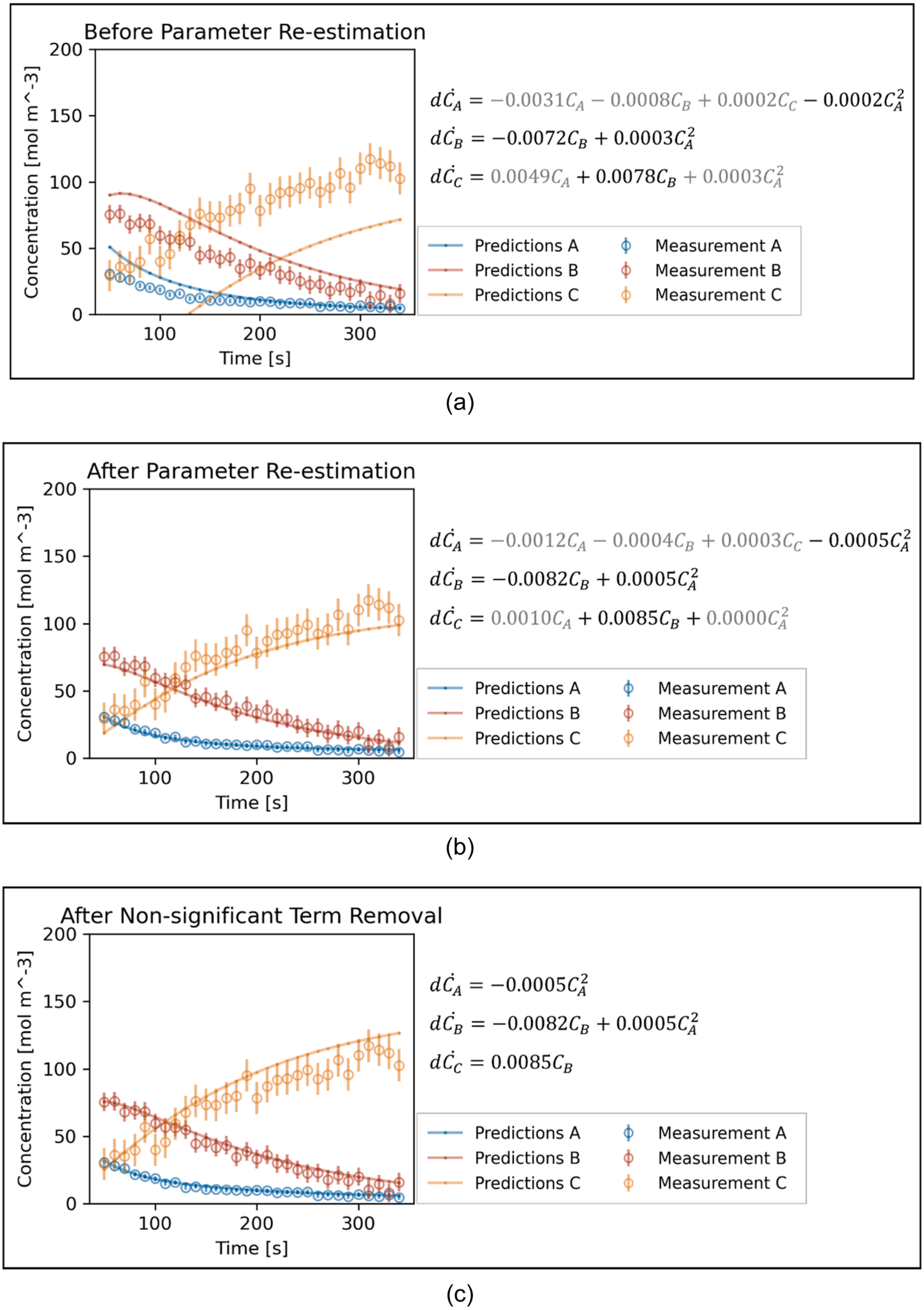


Fig. 3. Example of a generated model before and after parameter re-estimation and non-significant terms removal under conditions $N_{samples} = 30$, $\sigma = 10\%$, stopping criterion='normality', and experimental budget=15.

Regardless of whether any model meets the user-defined criterion, models are ranked by the Akaike Information Criterion (AIC) (Akaike, 1974) in ascending order to balance simplicity and accuracy. If one or more models satisfy the criterion, only those passing models are ranked. Otherwise, if no model meets the criterion and the experimental budget is exhausted, the models from the final iteration are ranked instead.

2.2. *Assessment criteria for models identified by DoE-SINDy*

The performance of models identified using the DoE-SINDy framework is assessed based on three primary criteria, statistically acceptable, structurally promising and structurally ground-truth.

The two structure-based criteria, structurally promising and structurally ground-truth, are applied only in-silico case studies, where the ground-truth model used to generate synthetic data is known. These criteria are used to evaluate whether DoE-SINDy can successfully recover the correct model structure. Regarding real experiments, where the true model is unknown, only the statistically acceptable criterion can be used for model evaluation.

- Statistically acceptable: A model is considered statistically acceptable if it passes a goodness-of-fit test, which evaluates whether the model sufficiently represents experimental data while accounting for measurement noise.
- Structurally promising: A model is structurally promising if it includes all terms from the ground-truth model but also contains additional terms. This is mathematically defined in Eq. (13):

$$\mathscr{I}(\boldsymbol{\theta}^{\boldsymbol{true}}) \subset \mathscr{I}(\widehat{\boldsymbol{\theta}}) \tag{13}$$

$$\mathscr{I}(\boldsymbol{\theta}^{\boldsymbol{true}}) = \mathscr{I}(\widehat{\boldsymbol{\theta}}) \tag{14}$$

- Structurally ground-truth: A model is structurally ground-truth if its structure exactly matches that of the ground-truth model, even if parameter estimates differ, as is shown in Eq. (14).

Here, $\boldsymbol{\theta}^{\boldsymbol{true}}$ and $\widehat{\boldsymbol{\theta}}$ denote the coefficient matrices of the ground-truth and identified models, respectively. The dimension of each matrix corresponds to the number of equations and the size of the candidate term library. Nonzero entries in these matrices indicate the inclusion of specific terms in the model, with their values representing the estimated parameters. The function $\mathscr{I}(\cdot)$ extracts the position indices of nonzero entries in the coefficient matrix as a set, so that the matches of the models are checked by comparing their set of position indices.

For instance, if the ground-truth coefficient matrix for a component A is $\theta_{c_A}^{true} = [2\ 0\ 0\ 3\ 0\ 0]$ and the identified coefficient matrix is $\widehat{\theta}_{c_A} = [2\ 0\ 0\ 3\ 1\ 0]$, the position indices sets are $\mathscr{I}\left(\theta_{c_A}^{true}\right) = \{1,\ 4\}$ and $\mathscr{I}(\widehat{\theta}_{c_A}) = \{1,\ 4,\ 5\}$. In this case, the model is structurally promising because $\{1,\ 4\} \subset \{1,\ 4,\ 5\}$, indicating that all ground-truth terms are included, albeit with additional terms. If the coefficient matrix of the identified model is $\widehat{\theta}_{c_A} = [1.8\ 0\ 0\ 3.2\ 0\ 0]$, $\mathscr{I}(\widehat{\theta}_{c_A}) = \{1,\ 4\} = \mathscr{I}\left(\theta_{c_A}^{true}\right)$. Thus, this model is considered structurally ground-truth as it contains all ground-truth terms and no additional terms.

Based on the criteria defined for assessing model performance, identified models are assigned labels summarised in Table 1. The most desirable outcome is the "TTT" scenario, where the model is both statistically acceptable and structurally matches the ground-truth. The "TTF" scenario is also acceptable, as the identified model contains all ground-truth terms, and additional terms can be refined through further experiments and parameter re-estimation.

In cases where the ground-truth model is unknown, models classified as "TFF" are selected based on their statistical adequacy. However, such models may later be rejected after model discrimination (Asprey and Macchietto, 2000). Scenarios such as "FTT," "FTF," and "FFF" arise when no acceptable models are identified within the available experimental budget. In these cases, adjustments to the candidate term library or modifications to the DoE-SINDy framework settings are recommended before conducting further experiments.

Table 1
Potential scenarios for the models identified by DoE-SINDy.

Performance	Statistically acceptable	Structurally promising	Structurally ground-truth	Circumstance
Best	T	T	T	Statistically acceptable and structurally ground-truth model identified
Good	T	T	F	Statistically acceptable and structurally promising model identified
Poor	T	F	F	Statistically acceptable but missing term(s) in the ground-truth model
	F	T	T	Structurally ground-truth but poor fit
	F	T	F	Structurally promising but poor fit
Worst	F	F	F	Poor fit and missing terms(s) in the ground-truth model

2.3. *Evaluation of DoE-SINDy*

A metric, ***Target Scenario Achievement Rate*** (***TSAR***), is introduced to evaluate the performance of DoE-SINDy in identifying models that achieve specific target scenarios. TSAR is calculated as the percentage of tests that successfully produce models meeting the target scenario, relative to the total number of tests analysed for the impact of these factors, as expressed in Eq. (12):

$$\mathrm{TSAR} = \frac{\text{Number of Tests with Models Meeting the Target Scenario}}{\text{Total Number of Tests}} \tag{15}$$

This metric is used to assess the impact of experimental design factors (e.g., initial concentrations, sample size, and experimental budget) and data conditions (e.g., noise level) on model identification outcomes.

3. Case study and implementation

We evaluate the performance of DoE-SINDy in recovering an assumed ground-truth kinetic model of a batch reaction system from in-silico data, using SINDy and ESINDy as benchmarks within the case study.

3.1. *Generation of in-silico data*

The considered chemical system is a three-component reacting mixture (A, B, C) reacted in an ideal mixed and isothermal batch reactor, following a series mechanism involving two reactions:

$$\mathrm{A} \xrightarrow{r_1} \mathrm{B} \ \ \mathrm{B} \xrightarrow{r_2} \mathrm{C} \tag{16}$$

where r_1 and r_2 represents the reaction rates in units of $\left[\mathrm{mol\ m^{-3}s^{-1}}\right]$. The reaction rate model consists of three ordinary differential equations representing concentration changes over time, with second-order kinetics for A and first-order for B.

$$d\dot{C}_A = -k_1 C_{\mathrm{A}}^2 = \theta_1 C_{\mathrm{A}}^2 d\dot{C}_B = k_1 C_{\mathrm{A}}^2 - k_2 C_{\mathrm{B}} = \theta_2 C_{\mathrm{A}}^2 + \theta_3 C_{\mathrm{B}} d\dot{C}_C = k_2 C_{\mathrm{B}} = \theta_4 C_{\mathrm{B}} \tag{17}$$

The dataset was generated through in-silico experiments, which is the same simulated case study used in Quaglio et al. (2020b). Experiments were conducted at 667 K, with rate constant $k_1 = 5 \times 10^{-4}$ $[\mathrm{mol}^{-1}\mathrm{m}^3\mathrm{s}^{-1}]$ and $k_2 = 7.8 \times 10^{-3}$ $[\mathrm{s}^{-1}]$. Each experimental run lasted 350 s, and the measurements were recorded every 10 s, starting at 50 s and continuing through 350 s. For obtaining multiple trajectories of time-series concentrations for model identification, the initial concentration of the component A is manipulated within the range from 40 to 250 $[\mathrm{mol\,m}^{-3}]$. The initial concentrations for component B and C are 0, as they are products of reaction 1 and 2. The maximum experimental budget is 15. Thus, 15 sets of initial concentrations are designed using Latin Hypercube Sampling to ensure exploration of the experimentally feasible region of operating conditions. Noise-free simulated data is generated by solving Eq. (17) with given initial concentrations via ***scipy.integrate.solve_ivp*** (Virtanen et al., 2020). In-silico measurements were generated by adding to the noise-free data the measurement noise, here characterised by a Gaussian distribution with zero mean and constant variance, assuming that the standard deviation is 10 % of the maximum values of concentrations when the initial concentrations of A is specified at 165 $[\mathrm{mol\,m}^{-3}]$. Thus, the standard deviations of the noise are approximately $[1.759,\ 6.899,\ 11.973][\mathrm{mol\,m}^{-3}]$. Kalman derivatives were implemented for numerically approximating the time derivatives of noise-added sampling points of the concentrations along the trajectory (Kaptanoglu et al., 2022). Because this study uses synthetic data, independent Gaussian noise with known variance is added. In real experiments, measurement errors can be heteroscedastic, non-Gaussian, auto-correlated, or affected by outliers. These situations will require a preprocessing stage on the raw data, such as variance-model estimation, robust filtering, and outlier rejection, before applying DoE-SINDy.

3.2. Model identification using DoE-SINDy

Model identification was performed with simulated state variables, including the concentrations of components A, B, and C, as well as their time derivatives. The initial iteration included six experiments, with up to 15 experiments conducted if no statistically acceptable model was identified. In each iteration, one new experiment was added, and the expanded dataset was randomly divided into training (80 %) and validation (20 %) sets, rounding up the training set size to the nearest integer if necessary.

The candidate term library for model generation comprised features: $\boldsymbol{g}(\boldsymbol{C}) = [C_{\mathrm{A}},\ C_{\mathrm{B}},\ C_{\mathrm{C}},\ C_{\mathrm{A}}^2,\ C_{\mathrm{B}}^2,\ C_{\mathrm{C}}^2]$. DoE-SINDy and SINDy both employed the sequentially thresholded least squares (STLSQ) algorithm for model generation, as implemented in ***PySINDy (v1.7.5)*** (Boninsegna et al., 2018; Kaptanoglu et al., 2022), using identical settings, including a threshold of 10^{-4}, matching the minimum parameter magnitude. In terms of ESINDy, we utilised a two-step ensemble method for model generation, also embedded in ***PySINDy (v1.7.5)***. Initially, ESINDy created a library ensemble to generate 1000 candidate models and calculated the inclusion probabilities of the library terms. Terms with inclusion probabilities below a predefined threshold, of 50 % were excluded. Next, the method employed the standard bagging ESINDy with STLSQ on the reduced library to generate another set of 1000 candidate models, taking the median of the identified coefficients to ensure robustness.

In step 4, for all three methods, models were ranked and preliminarily selected based on simplicity, with the top three simplest models retained for further evaluation. In model validation and selection step, the impact of four distinct stopping criteria, ‘and’, ‘chi2’, ‘normality’ and ‘or’, on model identification performance was assessed.

All computations were carried out under an Intel(R) Xeon(R) Gold 6140 CPU (36 cores, 2.30 GHz) running Red Hat Enterprise Linux 7.9, with 5 GB RAM available per job, under Python 3.9.

4. Results and discussion

This section presents an evaluation and comparison of the performance of original SINDy, ESINDy and DoE-SINDy in identifying kinetic models for a simulated batch reaction system. The analysis focuses on the accuracy of the identified models from the perspective of statistical accuracy and structural correctness, and effectiveness of original SINDy, ESINDy and DoE-SINDy in recovering the ground-truth model from the same noisy and small dataset. The results highlight the advantages of DoE-SINDy in delivering robust and simple models with fewer experiments.

4.1. Performance comparison of models identified by SINDy, ESINDy, and DoE-SINDy

The identification process begins with an initial dataset comprising 6 experiments, and the total experimental budget allows for up to 15 experiments. In each iteration, SINDy, ESINDy and DoE-SINDy are employed to identify a kinetic model from the same dataset. This iterative process continues until the predefined stopping criterion, ‘chi2’, is met. Table 2 summarises the iterative model identification process using these three approaches.

The row with the title N_{exp} represents the total number of experiments used in each iteration. For all three approaches, the same dataset is used when N_{exp} is identical. This table indicates the performance of the model identified in each iteration, using the labels defined in Table 1, which represents the statistical adequacy and structural correctness. As shown in Table 2, only DoE-SINDy successfully identified the model labelled ‘TTT’ among these three approaches, with a structure that matches the ground-truth and χ^2 test passed.

In terms of Table 3, despite slight differences between estimated parameters $\widehat{\boldsymbol{\theta}}$ from the reference $\boldsymbol{\theta}_{i,\mathbf{ref}}$, from its statistical adequacy point of view, the identified model fits the observations relatively well. The t values of the parameter estimates are significantly smaller than the reference, indicating substantial uncertainty in the parameter estimates. This uncertainty can be attributed primarily to parameter redundancy, which, in that the paired parameters should be combined, leads to unidentifiability. A secondary factor is the lack of sufficiently informative experimental data, which limits the ability to precisely estimate the parameters. This result highlights the need to integrate MBDoE into the framework to enhance parameter precision.

From the perspective of its graphical fit, the DoE-SINDy model matches the observations so well that the residuals are within the standard deviation of noise and follow a nearly Gaussian distribution. Thus, the model identified by DoE-SINDy is suitable for representing this system from both statistical and graphical point of view.

However, SINDy ended up with finding an ‘FFF’ model and ESINDy finding an ‘FTF’ model. The model structure of the SINDy identified model is very complex, and it does not contain the terms that are present in the ground-truth model, and the predictions are significantly deviate from the observations as shown in Fig. 4(a), indicating that the original SINDy performs poorly when the data is sparse and noisy. ESINDy works better than SINDy as it contains the ground-truth terms, but additional terms cause deviations in the profiles for component B and C as shown in Fig. 4(b). The results indicate that, although ESINDy performs better than SINDy, its performance is still insufficient to identify a suitable model for this system, especially when the data is noisy, sparse and limited.

Additionally, DoE-SINDy identified ‘TTT’ model in the 6th iteration with only 11 experiments. A structural promising model (‘FTF’) was identified using 8 experiments in the 3rd iteration, with model

Table 2

Model scenarios identified by original SINDy, ESINDy and DoE-SINDy across iterative experiment additions ($N_{samples}$ = 30, σ = 10%, stopping criterion='chi2', budget=15 experiments).

Iteration	1	2	3	4	5	6	7	8	9	10
$N_{\mathbf{exp}}$	6	7	8	9	10	11	12	13	14	15
Original SINDy	FFF	FTF	FTF	FTF	FFF	FTF	FTF	FTF	FTF	FFF
ESINDy	FTF	FTF	FTF	FTF	FTF	FTF	FTF	FTF	FTF	FTF
DoE-SINDy	-[i]	FFF	FTF	FTT	FTT	TTT[ii]	-	–	–	–

i. No identifiable model found.

ii. Model identification process stops at the 6th iteration as the model that meets the stopping criteria is identified.

Table 3

Estimated coefficients ($\widehat{\boldsymbol{\theta}}$) and corresponding t-values of models achieving the 'TTT' scenario under the stopping criterion 'chi2'. The training dataset has a noise level of $\sigma = 10\%$, sample size of 30 per experiment with a sampling interval of 10 s, and a total experimental budget of 15 experiments.

	$\boldsymbol{\theta}_{i,\mathbf{ref}}\ (\text{mol m}^{-3})$	$\widehat{\boldsymbol{\theta}}_i\ (\text{mol m}^{-3})$	$t(t_{\text{ref}}(95\%) = 1.97)$
θ_1	-5.00×10^{-4}	-4.51×10^{-4}	8.47×10^{-7}
θ_2	-7.80×10^{-3}	-8.18×10^{-3}	1.18×10^{-5}
θ_3	5.00×10^{-4}	4.94×10^{-4}	8.33×10^{-7}
θ_4	7.80×10^{-3}	8.48×10^{-3}	1.23×10^{-5}

complexity decreasing until the ground-truth model structure ('FTT') was reached in the 4th iteration. Example concentration profiles and corresponding equations of these 'FTF' and 'FTT' models are displayed in the Supplementary Figure 1 and Supplementary Table 1. Once the ground-truth model structure is identified, the identified structure remains at a low complexity level even when additional runs are used. In terms of statistically adequacy, DoE-SINDy successfully identified a model with ground-truth model structure and accurate fitting performance (model labelled 'TTT'). The performance of the SINDy models fluctuations over iterations. An 'FFF' model is identified in the 1st, 5th and 10th iterations; while a 'FTF' model is identified in the other iterations, with no evident trend of convergence. In terms of ESINDy, the models identified over iterations are consistently with promising structures, but the models are always much more complex than the ground-truth. DoE-SINDy has a higher likelihood of identifying a suitable model compared with the other two methods.

The comparison in Fig. 5 illustrates that DoE-SINDy identifies the ground-truth model more quickly (i.e. with a lower number of experimental runs) than the other two methods. Additionally, as the data size increases, the complexity of the models identified by DoE-SINDy incrementally decreases, eventually stabilising. This convergence suggests that the optimal and simplest governing model has likely been found. In contrast, neither SINDy nor ESINDy exhibit this trend. The number of parameters fluctuates irregularly and does not converge to a specific value. Even if the models identified by these methods meet the stopping criterion, such as goodness-of-fit test, we cannot confidently conclude that they represent the simplest and most suitable model for the system.

Notably, in the first iteration, the model identified by DoE-SINDy is not identifiable and is therefore rejected. Additionally, a complex model is more likely to include unidentifiable parameters. By integrating identifiability analysis increase the likelihood of rejecting complex models. In comparison, SINDy and ESINDy do not include an identifiability analysis step. Thus, SINDy and ESINDy is likely to identify a unidentifiable model, which renders them unreliable for accurately representing the system.

4.2. Success rate comparison for SINDy, ESINDy, and DoE-SINDy

We assessed the performance of original SINDy, ESINDy, and DoE-SINDy for recovering the ground-truth model under four different stopping criteria: 'and', 'chi2', 'normality' and 'or'. The model identification was iteratively conducted using identical 50 tests under a fixed experimental budget of 15 experiments, sampling size of 30 points with sampling interval of 10 s, and a standard deviation of 10 % of the maximum values of concentrations when the initial concentrations A is specified at 165 $\left[\text{mol m}^{-3}\right]$. Each approach's performance is quantified in terms of TSAR defined in Section 2.3, distinguishing between models meeting different success levels ('TTT', 'TTF', 'TFF', 'FTT', 'FTF' and 'FFF') based on statistical and structural adequacy defined in Section 2.2. Table 1. The TSAR (%) of original SINDy, ESINDy, and DoE-SINDy under four different stopping criteria is summarised in Table 4.

Original SINDy and ESINDy consistently failed to recover the ground-truth model across all tests and stopping criteria, as indicated by a TSAR of 0 % for 'TTT' and 'FTT'. Also, no statistically accurate model identified (TSAR of 0 % for 'TTT', 'TTF' and 'TFF'). Most identified models fell into the 'FTF' category (94 % for SINDy, 98 % for ESINDy), suggesting these models are structurally promising but fail goodness-of-fit tests because extra terms and inaccurate parameters introduce deviations.

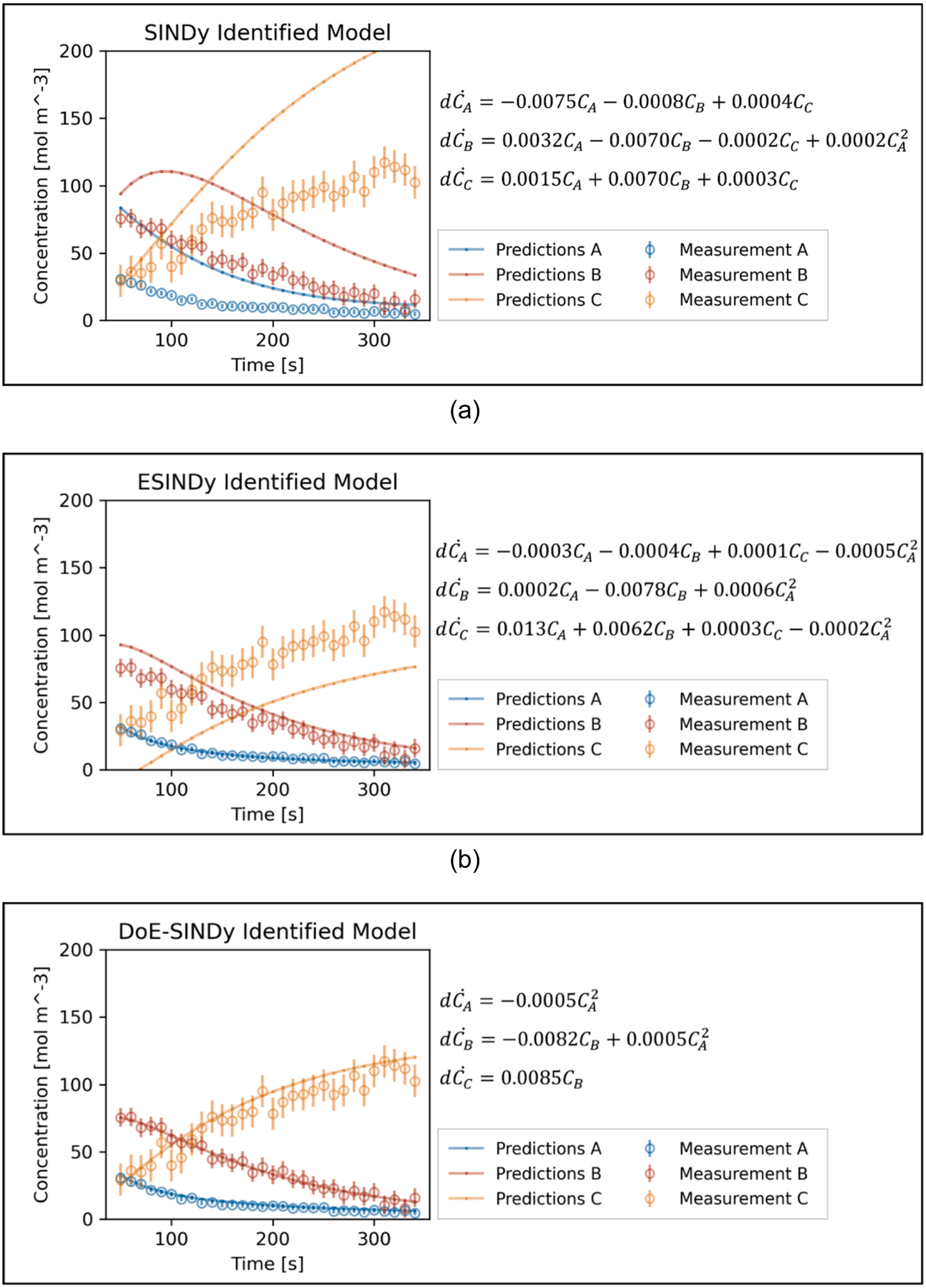
DoE-SINDy significantly outperformed the other two approaches, successfully identifying 'TTT' models under all stopping criteria. Performance depended on the criterion: the 'or' criterion achieved the highest TSAR for 'TTT' models (26 %), followed by the 'normality' criterion (22 %) and 'chi2' criterion (18 %). While the 'and' criterion resulted in the lowest (12 %). This indicates that the TSAR for 'TTT' models is inversely proportional to the restrictiveness of the stopping criteria—less restrictive criteria yield higher probabilities of identifying 'TTT' models. A similar trend was observed for 'TTF' models. Thus, criteria such as 'normality' and 'or' are more effective in identifying structurally accurate or promising models within a given dataset.

In practice, however, the ground-truth model structure is unknown, and statistical adequacy remains the primary criterion for model evaluation. The 'TFF' models introduces potential errors when applying these models in other feasible regions. When analysing statistically adequate models ('TXX'), the combined proportion of 'TTT' and 'TTF' under normality and or is approximately 70 %, whereas stricter criteria such as 'and' and 'chi2' achieve a higher combined proportion of 85 %, indicating greater reliability.

Nevertheless, stricter criteria exclude nearly 40 % of structurally promising or ground-truth models because of statistical inadequacy. Moreover, 38 % ('and') and 56 % ('chi2') of tests result in no identified suitable model when the experimental budget is exhausted. This indicates that convergence to a specific model structure becomes harder when stricter criteria are applied.

Table 5 provides the average computational times for SINDy, ESINDy, and DoE-SINDy under the four stopping criteria. The results indicate that DoE-SINDy has far higher computational cost compared to SINDy and ESINDy due to its iterative, weighted least squares parameter re-estimation. Specifically, when using stricter stopping criteria ('and' and 'chi2'), DoE-SINDy required over four times the runtime of the less strict criteria ('normality' and 'or'). This disparity is attributed to the larger number of iterations needed to satisfy stricter criteria. As additional experiments are incorporated, the number of candidate models also grows, slowing the process further.

In practice, stricter criteria ensure higher model reliability but come



(a)

(b)

(c)

Fig. 4. Predicted concentration profiles of the models identified by (a) original SINDy, (b) ESINDy, and (c) DoE-SINDy, compared with measurements ($\sigma = 10\%$, 30 samples per experiment, 10 s interval). Identified equations are displayed beside each plot. The ground-truth equations are given in Eq. (17): $d\dot{C}_A = -0.0005C_A^2$, $d\dot{C}_B = -0.0078C_B + 0.0005C_A^2$, $d\dot{C}_C = 0.0078C_B$.

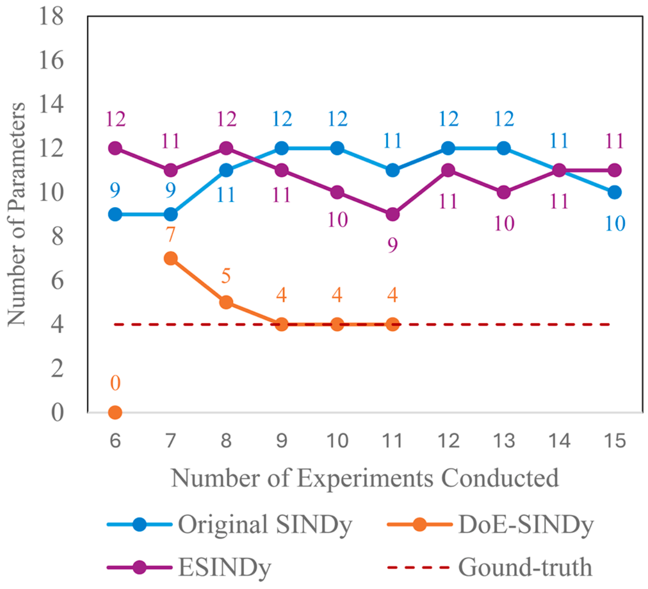


Fig. 5. Change in the number of parameters of the model identified by SINDy, ESINDy and the model ranked 1 identified by DoE-SINDy with iterative experiment additions.

Table 4
TSAR (%) of original SINDy, ESINDy and DoE-SINDy under four different stopping criteria.

Approach	Stopping Criteria	TTT	TTF	TFF	FTT	FTF	FFF
Original SINDy	and	0	0	0	0	94	6
	chi2	0	0	0	0	94	6
	normality	0	0	0	0	94	6
	or	0	0	0	0	94	6
ESINDy	and	0	0	0	0	98	2
	chi2	0	0	0	0	98	2
	normality	0	0	0	0	98	2
	or	0	0	0	0	98	2
DoE-SINDy	and	12	20	6	8	30	24
	chi2	18	30	8	8	26	10
	normality	22	42	32	0	4	0
	or	26	42	28	0	4	0

Table 5
Average runtime required by SINDy, ESINDy and DoE-SINDy to reach convergence under the stopping criteria ‘and’, ‘chi2’, ‘normality’ and ‘or’.

Stopping Criteria	and	chi2	normality	or
SINDy [min]	0.18	0.18	0.18	0.18
ESINDy [min]	1.37	1.32	1.33	1.32
DoE-SINDy [h]	2.39	2.09	0.52	0.50

at a computational cost. Less strict criteria reduce the number of experiments and runtime. Even though less strict criteria may yield a less reliable model (such as ‘TTF’ or ‘TFF’), they still provide a useful first approximation when time and budget is limited.

Of the four criteria, ‘normality’ is particularly promising, as it strikes a balance between identifying a high proportion of ‘TTT’ models, achieving a good combined proportion of ‘TTT’ and ‘TTF’, and maintaining reasonable identification speed.

5. Conclusion and future work

We propose DoE-SINDy, a design of experiments-integrated SINDy, for identifying kinetic model structures. DoE-SINDy outperforms existing methods by effectively addressing the challenges of limited experimental budgets, small datasets, and noise through an iterative framework integrating identifiability analysis, parameter re-estimation, structure simplification, and rigorous validation steps, ensuring statistically accurate, interpretable, and generalisable models with reduced complexity and accelerated convergence.

Three model generation approaches have been compared in this study: original SINDy, ESINDy and the proposed DoE-SINDy. Among the three approaches, DoE-SINDy is the only method capable of reliably identifying the ground-truth model (’TTT’) within the constraints of limited experimental budget and small dataset sizes. The iterative framework provides a clear converging trend, reducing model complexity and achieving convergence to the optimal model structure as the experimental dataset grows, which highlights DoE-SINDy’s ability to address the high variability issue in identified structures caused by using different training sets in existing methods. DoE-SINDy enhances robustness to noise by integrating experimental-level subsampling technique in the model generation step, reducing the inclusion of biased experiments that lead to overly complex models and ensuring a more representative and interpretable model structure. The integration of parameter re-estimation enhances the noise robustness and accuracy of the model on a broader region compared to the one obtained from generation step, as it is calibrated with full training set. The step for the removal of non-significant terms furtherly reduces model complexity, which ensures the generalisation of the identified model. The integration of a first identifiability analysis step in DoE-SINDy allows to reject unidentifiable and overly complex models before the computationally intensive parameter re-estimation step, accelerating the convergence process. A second identifiability check ensures model reliability after parameter re-estimation and non-significant model removal. By employing flexible optional stopping criteria, such as ‘normality’, DoE-SINDy balances computational efficiency with the success rate of identifying ground-truth models, addressing statistical accuracy without inflating runtime or costs. DoE-SINDy incorporates rigorous evaluation and AIC-based selection steps, balancing the statistical accuracy and model complexity of the finally confirmed model.

Despite the promising performance of the DoE-SINDy framework on identifying model structures, several challenges warrant further investigation:

- **Parameter redundancy**. Some coefficients that describe the same kinetic contribution remain separate, leading to redundancy and local unidentifiability. Future work will incorporate a rigorous procedure to detect these coefficients and re-parameterise the model accordingly.
- **Uncertainty in parameter estimates**. High uncertainty arises from both redundancy and limited information in the current data. In addition to the new re-parameterisation module, we will adopt model-based design of experiments for parameter precision (MBDoE-PP) to systematically design highly informative experiments and improve parameter precision efficiently.
- **Reducing runtime**. The computational cost of DoE-SINDy remains high due to the iterative nature of the approach and the integration of weighted regression. This can be mitigated by adopting more efficient parameter estimation algorithms, reducing the frequency of optimisation runs, or improving the speed of the optimisation process itself.
- **Structural uncertainty and model discrimination**. A model may fit statistically yet be structurally wrong. Beyond model complexity, additional metrics are needed to evaluate the structural adequacy of models. To address this, future work could incorporate MBDoE for model discrimination, which enables the exploration of conditions that better differentiate the performance of candidate models, enhancing the feasibility and reliability of the identified structures. Ideally, this approach should further reduce the number of experiments demanded.

- **Real-world application**. The present study assumes Gaussian, constant-variance noise. In future work we will (a) add an automated module for noise filtering and outlier rejection to preprocess raw experimental measurements, and (b) develop a noise-variance–model identification step that provides data-dependent weights for parameter re-estimation, thereby extending DoE-SINDy to noisier, heteroscedastic, and non-Gaussian datasets.

CRediT authorship contribution statement

Wenyao Lyu: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Federico Galvanin:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the funding from UCL EPSRC DTP Research Studentship (EP/R513143/1 and EP/W524335/1).

The current implementation of the DoE-SINDy framework is available from the corresponding author upon reasonable request. A fully documented open-source release will be deposited in a public repository after completion of the follow-up study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2025.109265.

Data availability

Data will be made available on request.

References

Abdullah, F., Alhajeri, M.S., Christofides, P.D., 2022a. Modeling and control of nonlinear processes using sparse identification: using dropout to handle noisy data. Ind. Eng. Chem. Res. 61, 17976–17992. https://doi.org/10.1021/acs.iecr.2c02639.

Abdullah, F., Wu, Z., Christofides, P.D., 2022b. Handling noisy data in sparse model identification using subsampling and co-teaching. Comput. Chem. Eng. 157, 107628. https://doi.org/10.1016/j.compchemeng.2021.107628.

Ahnert, K., Abel, M., 2007. Numerical differentiation of experimental data: local versus global methods. Comput. Phys. Commun. 177, 764–774. https://doi.org/10.1016/j.cpc.2007.03.009.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19, 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Angelis, D., Sofos, F., Karakasidis, T.E., 2023. Artificial intelligence in physical sciences: symbolic regression trends and perspectives. Arch. Comput. Methods Eng. 30, 3845–3865. https://doi.org/10.1007/s11831-023-09922-z.

Asprey, S.P., Macchietto, S., 2000. Statistical tools for optimal dynamic model building. Comput. Chem. Eng. 24, 1261–1267. https://doi.org/10.1016/S0098-1354(00)00328-8.

Asprey, S.P., Macchietto, S., Pantelides, C.C., 2000. Robust optimal designs for dynamic experiments. IFAC Proc. Vol. 33, 845–850. https://doi.org/10.1016/S1474-6670(17)38645-7.

Bard, Yonathan., 1974. Nonlinear Parameter Estimation. Academic Press, INC., New York.

Barz, T., López Cárdenas, D.C., Arellano-Garcia, H., Wozny, G., 2013. Experimental evaluation of an approach to online redesign of experiments for parameter determination. AIChE J 59, 1981–1995. https://doi.org/10.1002/aic.13957.

Bawa, S.G., Pankajakshan, A., Waldron, C., Cao, E., Galvanin, F., Gavriilidis, A., 2023. Rapid screening of kinetic models for methane total oxidation using an automated gas phase catalytic microreactor platform. Chem. Methods 3, e202200049. https://doi.org/10.1002/cmtd.202200049.

Bhadriraju, B., Bangi, M.S.F., Narasingam, A., Kwon, J.S., 2020. Operable adaptive sparse identification of systems: application to chemical processes. AIChE J 66, e16980. https://doi.org/10.1002/aic.16980.

Binns, M., Usai, A., Theodoropoulos, C., 2024. Identifiability methods for biological systems: determining subsets of parameters through sensitivity analysis, penalty-based optimisation, profile likelihood and LASSO model reduction. Comput. Chem. Eng. 186. https://doi.org/10.1016/j.compchemeng.2024.108683, 108683–108683.

Boninsegna, L., Nüske, F., Clementi, C., 2018. Sparse learning of stochastic dynamical equations. J. Chem. Phys. 148, 241723. https://doi.org/10.1063/1.5018409.

Bottmer, L., Croux, C., Wilms, I., 2022. Sparse regression for large data sets with outliers. Eur. J. Oper. Res. 297, 782–794. https://doi.org/10.1016/j.ejor.2021.05.049.

Brunton, S.L., Proctor, J.L., Kutz, J.N., Bialek, W., 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. U. S. A. 113, 3932–3937. https://doi.org/10.1073/pnas.1517384113.

Buzzi-Ferraris, G., Forzatti, P., 1983. A new sequential experimental design procedure for discriminating among rival models. Chem. Eng. Sci. 38, 225–232. https://doi.org/10.1016/0009-2509(83)85004-0.

Catchpole, E.A., Morgan, B.J.T., 1997. Detecting parameter redundancy. Biometrika 84, 187–196. https://doi.org/10.1093/biomet/84.1.187.

Chakraborty, A., Sivaram, A., Venkatasubramanian, V., 2021. AI-DARWIN: a first principles-based model discovery engine using machine learning. Comput. Chem. Eng. 154, 107470. https://doi.org/10.1016/j.compchemeng.2021.107470.

Chartrand, R., 2011. Numerical differentiation of noisy, nonsmooth data. ISRN Appl. Math. 2011, 1–11. https://doi.org/10.5402/2011/164564.

Chen, Y., Ierapetritou, M., 2020. A framework of hybrid model development with identification of plant-model mismatch. AIChE J 66, e16996. https://doi.org/10.1002/aic.16996.

Chen, Z., Liu, Y., Sun, H., 2021. Physics-informed learning of governing equations from scarce data. Nat. Commun. 12, 6136. https://doi.org/10.1038/s41467-021-26434-1.

D'Agostino, R., Stephens, M.A., 1986. Goodness-of-Fit-Techniques, Statistics: a Series of Textbooks and Monographs. CRC Press, New York.

Deussen, P., Galvanin, F., 2022. A model-based experimental design approach to assess the identifiability of kinetic models of hydroxymethylfurfural hydrogenation in batch reaction systems. Chem. Eng. Res. Des. 178, 609–622. https://doi.org/10.1016/j.cherd.2021.12.028.

Dobre, S., Bastogne, T., Profeta, C., Barberi-Heyob, M., Richard, A., 2012. Limits of variance-based sensitivity analysis for non-identifiability testing in high dimensional dynamic models. Automatica 48, 2740–2749. https://doi.org/10.1016/j.automatica.2012.05.004.

Draper, N.R., Smith, H., 1998. Applied regression analysis. Wiley Series in Probability and Statistics, 3rd ed. Wiley, New York.

Edwards, K., Edgar, T.F., Manousiouthakis, V.I., 2000. Reaction mechanism simplification using mixed-integer nonlinear programming. Comput. Chem. Eng. 24, 67–79. https://doi.org/10.1016/S0098-1354(00)00311-2.

Fasel, U., Kutz, J.N., Brunton, B.W., Brunton, S.L., 2022. Ensemble-SINDy: robust sparse model discovery in the low-data, high-noise limit, with active learning and control. Proc. R. Soc. Math. Phys. Eng. Sci. 478, 20210904. https://doi.org/10.1098/rspa.2021.0904.

Franceschini, G., Macchietto, S., 2008. Model-based design of experiments for parameter precision: state of the art. Chem. Eng. Sci., Model-Based Experimental Analysis 63, 4846–4872. https://doi.org/10.1016/j.ces.2007.11.034.

Galvanin, F., Macchietto, S., Bezzo, F., 2007. Model-based design of parallel experiments. Ind. Eng. Chem. Res. 46, 871–882. https://doi.org/10.1021/ie0611406.

Gao, F., Han, L., 2012. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. Comput. Optim. Appl. 51, 259–277. https://doi.org/10.1007/s10589-010-9329-3.

Gottu Mukkula, A.R., Paulen, R., 2019. Optimal experiment design in nonlinear parameter estimation with exact confidence regions. J. Process Control 83, 187–195. https://doi.org/10.1016/j.jprocont.2019.01.004.

Hone, C.A., Boyd, A., O'Kearney-McMullan, A., Bourne, R.A., Muller, F.L., 2019. Definitive screening designs for multistep kinetic models in flow. React. Chem. Eng. 4, 1565–1570. https://doi.org/10.1039/C9RE00180H.

Hunter, W.G., Reiner, A.M., 1965. Designs for discriminating between two rival models. Technometrics 7, 307–323. https://doi.org/10.1080/00401706.1965.10490265.

Javaid, M., Haleem, A., Suman, R., 2023. Digital Twin applications toward Industry 4.0: a review. Cogn. Robot. 3, 71–92. https://doi.org/10.1016/j.cogr.2023.04.003.

Johansen, T.A., 1997. On Tikhonov regularization, bias and variance in nonlinear system identification. Automatica 33, 441–446. https://doi.org/10.1016/S0005-1098(96)00168-9.

Jul-Rasmussen, P., Chakraborty, A., Venkatasubramanian, V., Liang, X., Huusom, J.K., 2024. Hybrid AI modeling techniques for pilot scale bubble column aeration: a comparative study. Comput. Chem. Eng. 185, 108655. https://doi.org/10.1016/j.compchemeng.2024.108655.

Kaheman, K., Brunton, S.L., Kutz, J.N., 2022. Automatic differentiation to simultaneously identify nonlinear dynamics and extract noise probability distributions from data. Mach. Learn. Sci. Technol. 3, 015031. https://doi.org/10.1088/2632-2153/ac567a.

Kaheman, K., Kutz, J.N., Brunton, S.L., 2020. SINDy-PI: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics. Proc. R. Soc. Math. Phys. Eng. Sci. 476, 20200279. https://doi.org/10.1098/rspa.2020.0279.

Kaptanoglu, A.A., Callaham, J.L., Aravkin, A., Hansen, C.J., Brunton, S.L., 2021. Promoting global stability in data-driven models of quadratic nonlinear dynamics. Phys. Rev. Fluids 6, 094401. https://doi.org/10.1103/PhysRevFluids.6.094401.

Kaptanoglu, A.A., Silva, B.M.De, Fasel, U., Kaheman, K., Goldschmidt, A.J., Callaham, J., Delahunt, C.B., Nicolaou, Z.G., Champion, K., Loiseau, J.-C., Kutz, J.N., Brunton, S. L., 2022. PySINDy: a comprehensive Python package for robust sparse system identification. J. Open Source Softw. 7, 3994. https://doi.org/10.21105/joss.03994.

D.C. López, C., Barz, T., Körkel, S., Wozny, G., 2015. Nonlinear ill-posed problem analysis in model-based parameter estimation and experimental design Comput. Chem. Eng. 77, 24–42. https://doi.org/10.1016/j.compchemeng.2015.03.002.

Mandel, J., 1964. The Statistical Analysis of Experimental Data. Interscience, New York.

Mangan, N.M., Brunton, S.L., Proctor, J.L., Kutz, J.N., 2016. Inferring biological networks by sparse identification of nonlinear dynamics. IEEE Trans. Mol. Biol. Multi-Scale Commun. 2, 52–63. https://doi.org/10.1109/TMBMC.2016.2633265.

Maria, G., 2004. A review of algorithms and trends in kinetic model identification for chemical and biochemical systems. Chem. Biochem. Eng. Q. 18, 195–222.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245. https://doi.org/10.2307/1268522.

Mclean, K.A.P., Mcauley, K.B., 2012. Mathematical modelling of chemical processes—Obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. Can. J. Chem. Eng. 90, 351–366. https://doi.org/10.1002/CJCE.20660.

Meneghetti, N., Facco, P., Bezzo, F., Barolo, M., 2014. A methodology to diagnose process/model mismatch in first-principles models. Ind. Eng. Chem. Res. 53, 14002–14013. https://doi.org/10.1021/ie501812c.

Messenger, D.A., Bortz, D.M., 2021. Weak SINDy for partial differential equations. J. Comput. Phys. 443, 110525. https://doi.org/10.1016/j.jcp.2021.110525.

Miao, H., Xia, X., Perelson, A.S., Wu, H., 2011. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM Rev 53, 3–39. https://doi.org/10.1137/090757009.

Molga, E., Cherbański, R., 1999. Hybrid first-principle–neural-network approach to modelling of the liquid–liquid reacting system. Chem. Eng. Sci. 54, 2467–2473. https://doi.org/10.1016/S0009-2509(98)00506-5.

Narayanan, H., Cruz Bournazou, M.N., Guillén Gosálbez, G., Butté, A., 2022. Functional-hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. Chem. Eng. J. 430, 133032. https://doi.org/10.1016/j.cej.2021.133032.

Quaglio, M., Fraga, E.S., Galvanin, F., 2020a. A diagnostic procedure for improving the structure of approximated kinetic models. Comput. Chem. Eng. 133, 106659. https://doi.org/10.1016/j.compchemeng.2019.106659.

Quaglio, M., Roberts, L., Bin Jaapar, M.S., Fraga, E.S., Dua, V., Galvanin, F., 2020b. An artificial neural network approach to recognise kinetic models from experimental data. Comput. Chem. Eng. 135, 106759. https://doi.org/10.1016/j.compchemeng.2020.106759.

Quaglio, M., Waldron, C., Pankajakshan, A., Cao, E., Gavriilidis, A., Fraga, E.S., Galvanin, F., 2019. An online reparametrisation approach for robust parameter estimation in automated model identification platforms. Comput. Chem. Eng. 124, 270–284. https://doi.org/10.1016/j.compchemeng.2019.01.010.

Rosafalco, L., Conti, P., Manzoni, A., Mariani, S., Frangi, A., 2024. EKF–SINDy: empowering the extended Kalman filter with sparse identification of nonlinear dynamics. Comput. Methods Appl. Mech. Eng. 431, 117264. https://doi.org/10.1016/j.cma.2024.117264.

Sangoi, E., Cattani, F., Padia, F., Galvanin, F., 2025. Foliar uptake of biocides: statistical assessment of compartmental and diffusion-based models. Chem. Eng. Sci. 317, 121984. https://doi.org/10.1016/j.ces.2025.121984.

Sangoi, E., Quaglio, M., Bezzo, F., Galvanin, F., 2024. An optimal experimental design framework for fast kinetic model identification based on artificial neural networks. Comput. Chem. Eng. 187, 108752. https://doi.org/10.1016/j.compchemeng.2024.108752.

Schaber, S.D., Born, S.C., Jensen, K.F., Barton, P.I., 2014. Design, execution, and analysis of time-varying experiments for model discrimination and parameter estimation in microreactors. Org. Process Res. Dev. 18, 1461–1467. https://doi.org/10.1021/op500179r.

Schwaab, M., Silva, F.M., Queipo, C.A., Barreto, A.G., Nele, M., Pinto, J.C., 2006. A new approach for sequential experimental design for model discrimination. Chem. Eng. Sci. 61, 5791–5806. https://doi.org/10.1016/J.CES.2006.04.001.

Schweidtmann, A.M., Zhang, D., von Stosch, M., 2024. A review and perspective on hybrid modeling methodologies. Digit. Chem. Eng. 10, 100136. https://doi.org/10.1016/j.dche.2023.100136.

Sen, M., Arguelles, A.J., Stamatis, S.D., García-Muñoz, S., Kolis, S., 2021. An optimization-based model discrimination framework for selecting an appropriate reaction kinetic model structure during early phase pharmaceutical process development. React. Chem. Eng. 6, 2092–2103. https://doi.org/10.1039/D1RE00022H.

Silva, B.M.De, Champion, K., Quade, M., Loiseau, J.-C., Kutz, J.N., Brunton, S.L., 2020. PySINDy: a Python package for the sparse identification of nonlinear dynamical systems from data. J. Open Source Softw. 5, 2104. https://doi.org/10.21105/joss.02104.

Silvey, S.D., 1975. Statistical inference. Monographs On Statistics and Applied Probability, 2nd ed. Chapman and Hall, London.

Sjöberg, J., Ljung, L., 1992. Overtraining, regularization, and searching for minimum in neural networks. In: IFAC Proc. Vol., 4th IFAC Symposium on Adaptive Systems in Control and Signal Processing 1992. Grenoble, France, pp. 73–78. https://doi.org/10.1016/S1474-6670(17)50715-6, 1-3 July 25.

Sjöberg, J., McKelvey, T., Ljung, L., 1993. On the use of regularization in system identification. In: IFAC Proc. Vol., 12th Triennal Wold Congress of the International Federation of Automatic control. Volume 5 Associated Technologies and Recent Developments. Sydney, Australia, pp. 75–80. https://doi.org/10.1016/S1474-6670(17)48226-7, 18-23 July 26.

Tibshirani, R.J., Taylor, J., 2011. The solution path of the generalized lasso. Ann. Stat. 39, 1335–1371. https://doi.org/10.1214/11-AOS878.

Tikhonov, A.N., Arsenin, V.I., 1977. Solutions of Ill-Posed problems, Scripta series in Mathematics. Winston, Washington.

Tsay, C., Pattison, R.C., Baldea, M., Weinstein, B., Hodson, S.J., Johnson, R.D., 2017. A superstructure-based design of experiments framework for simultaneous domain-restricted model identification and parameter estimation. Comput. Chem. Eng. 107, 408–426. https://doi.org/10.1016/j.compchemeng.2017.02.014.

Vanrolleghem, P.A., Daele, M.V., Dochain, D., 1995. Practical identifiability of a biokinetic model of activated sludge respiration. Water Res 29, 2561–2570. https://doi.org/10.1016/0043-1354(95)00105-T.

Venkatasubramanian, V., 2019. The promise of artificial intelligence in chemical engineering: is it here, finally? Ai. Che. J 65, 466–478. https://doi.org/10.1002/aic.16489.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Meth. 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Waldron, C., Pankajakshan, A., Quaglio, M., Cao, E., Galvanin, F., Gavriilidis, A., 2019. Closed-loop model-based design of experiments for kinetic model discrimination and parameter estimation: benzoic acid esterification on a heterogeneous catalyst. Ind. Eng. Chem. Res. 58, 22165–22177. https://doi.org/10.1021/acs.iecr.9b04089.

Wei, B., 2022. Sparse dynamical system identification with simultaneous structural parameters and initial condition estimation. Chaos Solit. Fractals 165, 112866. https://doi.org/10.1016/j.chaos.2022.112866.

Zhang, D., Savage, T.R., Cho, B.A., 2020. Combining model structure identification and hybrid modelling for photo-production process predictive simulation and optimisation. Biotechnol. Bioeng. 117, 3356–3367. https://doi.org/10.1002/bit.27512.

Zheng, P., Askham, T., Brunton, S.L., Kutz, J.N., Aravkin, A.Y., 2019. A unified framework for sparse relaxed regularized regression: SR3. IEEE Access 7, 1404–1423. https://doi.org/10.1109/ACCESS.2018.2886528.