Systems & Control Transactions

Original Research Article
Peer Reviewed Conference Proceeding
ESCAPE 36 - European Symposium on Computer Aided Process Engineering
Sheffield, UK. 21-24 June 2026

PSE PRESS

Comprehensive Framework for Model Discovery and Discrimination Based on Symbolic Regression and Structural Identifiability - Application to a Partially Observed Chemical Reaction System

Xuming Yuan[a], Brahim Benyahia[a]

[a] Loughborough University, Department of Chemical Engineering, Loughborough, Leicestershire, UK
* Corresponding Author: B.Benyahia@lboro.ac.uk.

ABSTRACT

Traditional approaches for mechanistic modelling require in-depth understanding of the underlying chemical and physical phenomena to construct reliable and predictive models. However, at early stages of development, limited experimental data, incomplete expert knowledge, and non-observable states often hinder a full understanding of the underlying mechanisms. Symbolic regression (SR) enables systematic model discovery and offers a practical route to addressing these challenges by automating the identification of interpretable model structures and the estimation of associated parameters from available data. However, structural identifiability and observability (SIO), a critical property of such models, is often overlooked in SR, thereby limiting its broader adoption and effective deployment. To address these limitations, this study proposes a comprehensive framework, which leverages scarce prior knowledge in SR and incorporates SIO analysis, offering a potential solution to capture the effects of all state variables and ensure rigorous structure of the resulting models. The proposed strategy is demonstrated on a partially observed sulfide oxidation system, demonstrating how an SIO-assured model can be systematically identified from limited data and incomplete system knowledge. Overall, this work presents a potential pathway for extending model discovery to partially observed systems and enhancing model robustness, thereby positioning SR as an alternative tool with respect to traditional kinetic modeling strategies.

Keywords: Modelling and Simulations, Symbolic Regression, Partially Observed Systems, Structural Identifiability & Observability Analysis, Systematic Model Development

1. INTRODUCTION

The availability of robust mathematical models and digital twins constitutes a cornerstone of digital transformation across multiple industries [1, 2]. Mathematical models are essential for both off-line and real-time decision-making and underpin process design, optimization, scale-up, control, operation, and integration [3, 4]. The recent emergence of digitally driven paradigms for process and product quality and environmental sustainability, such as Green-by-Design and Quality-by-Digital-Design frameworks [5-7], has further reinforced the central role of mathematical models, particularly within the pharmaceutical industry. These recent shifts coincide with the broader adoption of continuous manufacturing technologies, collectively driving increased demand for high-quality mathematical models that support design space identification and exploration, start-up and shut down, and control strategies [8, 9]

Traditionally, the development of a mathematical model starts by postulating the inputs and outputs, then constructing the model equations which involve the unknown model parameters to be calibrated using experimental data. Often the underlying phenomena to be modelled are intricate, subject to limited measurements, non-observables, and expert bias. These challenges result in multiple model candidates which may require further benchmarking and discrimination procedures. However, the identification of affective model through the traditional route may demand extensive prior knowledge of

the system and domain expertise, which can be scarce especially for early-stage studies. More importantly, the optimality of the chosen kinetic expression is probably questionable since discrimination is normally restricted to a handful of manually crafted alternatives, while an exhaustive search remains prohibitively labor-intensive. On the contrary, data-driven models (e.g., response surface models, artificial neural networks) dispense with mechanistic assumptions and exhibit superior adaptability to the data supplied, offering accurate predictions within the training space. Nonetheless, they are known to be data-hungry and commonly difficult to extrapolate. In addition, they are operated as opaque black boxes with poor interpretability, which potentially hinders the comprehension of the underlying natural phenomenon.

The emergence of automated model discovery approaches is possibly delivering a way out of this impasse. Through automating the exploration and identification of the most appropriate model over a large combinatorial search space, the demand of human data-science expertise is minimized, thereby dramatically reduces the effort for the determination of the optimal set of model equations. In particular, symbolic regression (SR), a process that discovers parsimonious mathematical expressions from datasets, is a group of the most attractive approaches in automated model discovery. SR refers to the practice not only estimating the numerical coefficients of a pre-defined mathematical expression but also exploring the most appropriate functional forms describing the data [14], which fuses applied mathematics and machine learning to provide deep insights of the underlying relationship between the input, output and state variables with limited prerequisites of the model structure. Therefore, it is greatly beneficial to unveil the unknown governing equations of the systems that manifest complex or nonlinear behaviors. Hence, SR is rapidly becoming a catalyst for advanced model development. For instance, two automated model discovery frameworks (ADoK-S and ADoK-W) were developed for chemical reactions to automatically yield process models utilizing a Python-based deep SR toolbox PySR [11]. SR is also applied to discover the nucleation and growth expressions for a batch crystallization case of potassium sulphate, yielding highly resembling equations as the ones of standard method of moments, confirming the interpretability of the model and the effectiveness of the methodology [10]. Another SR implementation on crystallization process [15] leverages SINDy to obtain the nucleation and growth governing equations, which fitted well with the 1st and 2nd moments data but were not very satisfactory for 0th and 3rd moments, probably due to the invalidity of the inherent assumption of linear combination of the terms in SINDy. In a biological context, the kinetic model of a microbial fermentation was derived through implementing SR, which was then embedded in a multi-objective optimization to maximize cell biomass, protein titre and minimize batch time [10]. Experimental validation showed excellent agreement with model predictions, demonstrating the practical utility of SR in real-world bioprocesses.

SR is an extraordinary tool to uncover concise, predictive equations, yet most studies stop at goodness-of-fit and ignore structural identifiability and observability (SIO). Structural identifiability asks whether the parameters can be uniquely recovered from the observables, while observability determines whether unmeasured states can be uniquely inferred from the observables. Neglecting either property incurs ill-posed models that fit the training data yet yield unreliable predictions [17].

To bridge this gap, we couple PySR with the MATLAB-based toolbox STRIKE-GOLDD and its extension AutoRepar [16]. PySR is selected as the algorithm for model generation in this study due to greater flexibility in the functional forms and better compatibility to unmeasured states compared with SINDy. After PySR proposes a parsimonious model, STRIKE-GOLDD investigates it SIO, and the equations will be reparametrized by AutoRepar if any structural defect is pinpointed. The loop repeats until a structurally sound, data-efficient model is returned.

The resulting workflow furnishes an end-to-end pipeline from raw data to trustworthy equations, requiring minimal prior mechanistic insight and human intervention. We demonstrate its value on a partially observed sulfide-oxidation system [12], which starts from sparse measurements and recovers a compact, identifiable and observable high-fidelity model. By embedding SIO analysis inside SR, the approach offers a rigorous new paradigm for early-stage process development where data are scarce and mechanistic knowledge is limited.

2. METHODOLOGY

The proposed holistic framework which integrates the automated model discovery and SIO analysis for chemical reaction systems as shown in Figure 1. Initially, PySR is used to perform SR based on the measured data of the observables ($\boldsymbol{y}$), generating a large set of plausible concentration-time equations $\boldsymbol{y}(t)$. The $\boldsymbol{y}(t)$ with the lowest Akaike information criterion (AIC) is chosen, approximating the concentration evolution of the observed species. Thereafter, the concentration-time expressions for the entire set of state variables $\boldsymbol{x}(t)$ are derived from $\boldsymbol{y}(t)$ applying mass balance and stoichiometry, which allows the numerical computation of derivative values $\dot{\boldsymbol{x}}(t)$. Next, the evaluated $\dot{\boldsymbol{x}}(t)$ values are leveraged in couple with the measured data of $\boldsymbol{y}$ in the SR exploring $\dot{\boldsymbol{y}}(\boldsymbol{x})$ expressions using PySR. It is worth highlighting that $\dot{\boldsymbol{y}}(\boldsymbol{x})$ is the kinetic equations set describing the chemical reaction system. Afterwards, an AIC test is performed to find the top N

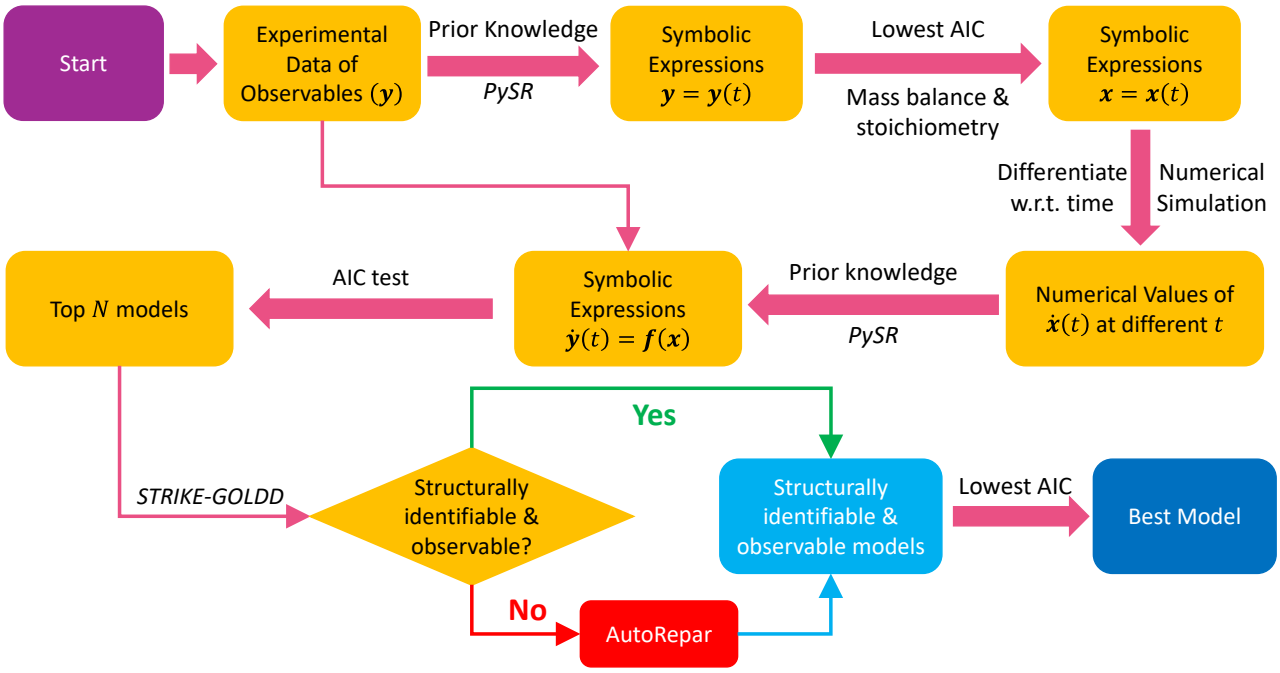

Figure 1. Proposed framework which integrates the automated model discovery and SIO analysis for chemical reaction systems.

models best fitting the data. The SIO of these models are scrutinized using STRIKE-GOLDD, while the models with structural defects will be reparametrized by AutoRepar. Eventually, the SIO-assured model with the lowest AIC is selected as the best model. This pipeline delivers models that both fit the data and exhibit rigorous structure, minimizing the requirement of prior mechanistic knowledge and domain expertise, thereby reducing human input to problem specification and data upload.

3. CASE STUDY

A sulfide oxidation reaction system [12] is selected as the case study, which involves four species: raw material methyl phenyl sulfide (R), desired product methyl phenyl sulfoxide (D), side-product methyl phenyl sulfone (U) and reactant hydrogen peroxide (H_2O_2). The main reaction and the side reaction are presented in Equations 1 and 2, respectively. An experimental study was conducted in [12], providing the concentration-time data of R, D and U, as depicted in Figure 2, while the initial concentration of H_2O_2 is 2.5 times as the initial concentration of R. The process involves four state variables $\boldsymbol{x} = [\mathrm{R}, \mathrm{D}, \mathrm{U}, H_2O_2]$ and three observables $\boldsymbol{y} = [\mathrm{R}, \mathrm{D}, \mathrm{U}]$. The goal is to find the underlying rate equations of $\mathrm{R}, \mathrm{D}, \mathrm{U}$ and H_2O_2 based on the data provided.

$$\mathrm{R} + H_2O_2 \rightarrow \mathrm{D} + H_2O \tag{1}$$

$$\mathrm{D} + H_2O_2 \rightarrow \mathrm{U} + H_2O \tag{2}$$

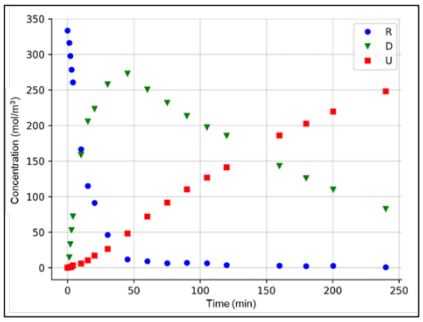

Figure 2. Experimental concentration-time data of methyl phenyl sulfide oxidation [12]. R: Methyl phenyl sulfide. D: Methyl phenyl sulfoxide. U: Methyl phenyl sulfone.

3.1. Obtain $\boldsymbol{y}(t), \boldsymbol{x}(t)$ and $\dot{\boldsymbol{x}}(t)$

The first step in the workflow is to obtain a concentration-time model for the observables denoted by $\boldsymbol{y}(t)$. Before running SR, it should be noticed that the atomic balance of sulfur can be given by Equation 3, where $\mathrm{R}_{\mathrm{ini}}$ the initial concentration of R. Since $\mathrm{R}_{\mathrm{ini}}$ is a known constant, U can be treated as a dependent variable, thereby it is only necessary to symbolically regress $\mathrm{R}(t)$ and $\mathrm{D}(t)$. The PySR settings in this step are summarized in Table 1.

$$\mathrm{R}(t) + \mathrm{D}(t) + \mathrm{U}(t) = \mathrm{R}_{\mathrm{ini}} \tag{3}$$

By running SR, a series of mathematical expressions of $\mathrm{R}(t)$ and $\mathrm{D}(t)$ are obtained. In this case, 19 candidates for $\mathrm{R}(t)$ and 17 candidates for $\mathrm{D}(t)$ are identified by PySR, among which the expressions corresponding to the

Table 1: PySR settings for the discovery of concentration-time relationships

Setting	Value	Description
niterations	100	Maximum number of iterations in symbolic regression
binary_operators	["+", "-", "*", "/"]	Allowed computational operations between two terms
unary_operators	["exp"]	Allowed computational operations on one term
loss	loss(x, y) = (x - y)^2	Fitting objective function: mean squared error
population_size	100	Population size for seeking the mathematical expression
maxsize	30	Complexity limit of the mathematical expression
timeout_in_seconds	60	Time limit of symbolic regression search
parsimony	10^{-3}	Parsimony coefficient determining the sparsity of the mathematical expression

Complexity	Loss	Equation
1	88.72346	-1.5131524
3	0.65134954	-0.063507065 * R
5	0.64904344	(-0.06375289 * R) + 0.06231785
7	0.42687345	R * (-0.061005794 + (-5.4023287e-5 * D))
9	0.14379378	(-0.057983927 + ((D * R) * -5.498032e-7)) * R
11	0.08519997	(-0.04215632 + (R * ((D * -8.229583e-7) + -4.6016456e-5))) * R
13	0.0780262	((-0.057921443 - ((R * 8.5146423e-10) * (D * H2O2))) * R) - -0.19413921
15	0.07224689	(R * (-0.057804443 - (((R - U) * 8.545725e-10) * (D * H2O2)))) - -0.13526249
17	0.055178992	(27.877422 + R) * (-0.05317548 + (((D * (-76.895355 + R)) - (H2O2 * U)) * -8.234853e-7))
19	0.055041444	(27.877422 + R) * (-0.05316254 + ((((D * (R + -76.895355)) - (H2O2 * U)) + U) * -8.2493364e-7))

(a)

Complexity	Loss	Equation
1	78.04503	1.3129557
3	1.8908302	R * 0.058173776
5	1.2114038	-1.0696173 + (R * 0.062388875)
7	1.2113937	((R * 1.0623888) - 1.0695953) - R
9	0.5922489	(R * ((D * 9.781079e-5) + 0.059605878)) - 1.5127325
11	0.11398397	((0.05509739 + (D * (8.210762e-7 * R))) * R) - 1.3124856
13	0.0565364	((0.060073934 + ((8.721774e-7 * R) * D)) * R) - (H2O2 * 0.0036195528)
15	0.0357263	(R * (0.05330363 + ((3.2462417e-6 * ((R * 0.2775878) - U)) * D))) - 0.83941805
17	0.03287169	((0.05330363 + ((3.7689076e-6 * ((R * 0.22400005) - (U + -3.4678288))) * D)) * R) - 0.8074039
19	0.031657755	(R * (0.05332232 + ((4.102189e-6 * ((0.194399 * R) - (U + (-2.7117753 - 2.7252011)))) * D))) - 0.780967

(b)

Figure 3. Symbolically regressed reaction rate expressions of: (a) Methyl phenyl sulphide (R), (b) Methyl phenyl sulfoxide (D).

lowest AIC values are selected as the best concentration-time equations (Equations 4 and 5 below).

$$\mathrm{R}(t) = \frac{\exp(8.678)}{t+17.577+0.0732(\exp(0.192t)+t^2)} + \exp(1.469) \quad (4)$$

$$\mathrm{D}(t) = \exp\left(\exp\left(0.645 + \frac{t+114.992}{t+\frac{212.903}{t+1.676}+95.674}\right)\right) - 0.559t - 24.232 \quad (5)$$

With $\mathrm{R}(t)$ and $\mathrm{D}(t)$ determined, the mathematical expression of $\mathrm{U}(t)$ can be readily calculated using Equation 3. To derive $\mathrm{H_2O_2}(t)$, the mass balance can be implemented, deducing that the consumption of $\mathrm{H_2O_2}$ is equal to the sum of the consumption of R and the generation of U, resulting in Equation 6. As the initial concentrations of R ($\mathrm{R_{ini}}$) and $\mathrm{H_2O_2}$ ($\mathrm{H_2O_{2_{ini}}}$) are known, Equation 6 enables the calculation of $\mathrm{H_2O_2}$ value at any given time.

$$\mathrm{H_2O_2}(t) = \mathrm{H_2O_{2_{ini}}} - \left(\mathrm{R_{ini}} - \mathrm{R}(t)\right) - \mathrm{U}(t) \quad (6)$$

Therefore, $\boldsymbol{x}(t) = [\mathrm{R}(t), \mathrm{D}(t), \mathrm{U}(t), \mathrm{H_2O_2}(t)]$ is obtained, which allows the numerical computation of the $\dot{\boldsymbol{x}}(t)$ values at the sampling points, approximating the reaction rates

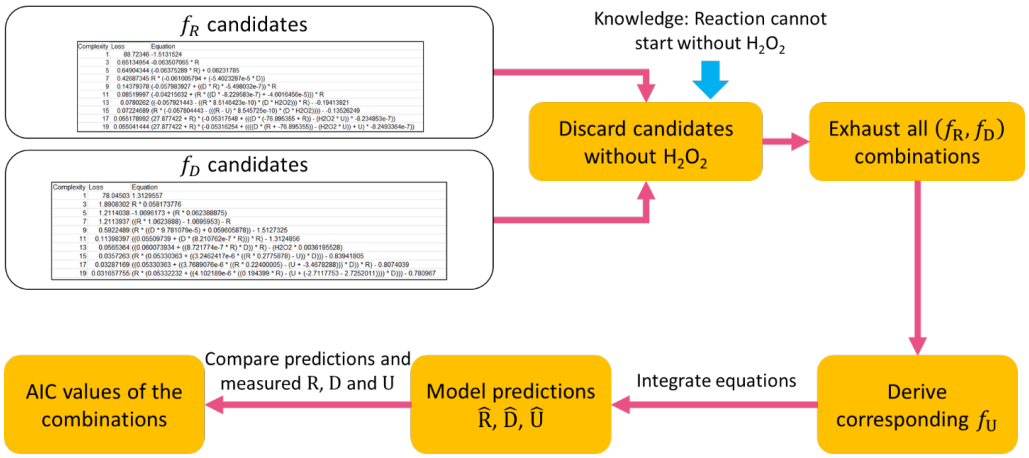

Figure 4. Procedure to construct valid rate model candidates and calculate the AIC values.

for each species in the experiment. The measured of $\mathrm{R, D, U}$ data, the simulated H_2O_2 values, together with the simulated $\dot{\boldsymbol{x}}(t)$ values at the sampling points, are subsequently used in the next step discovering the $\dot{\boldsymbol{y}}(\boldsymbol{x})$ mathematical expressions.

3.2. Derive $\dot{\boldsymbol{y}}(\boldsymbol{x})$

As emphasized in section 3.1, U and H_2O_2 are dependent on R and D, thus it is only necessary to attain the mathematical expressions $\dot{\mathrm{R}} = f_{\mathrm{R}}(\mathrm{R, D, U}, H_2O_2)$ and $\dot{\mathrm{D}} = f_{\mathrm{D}}(\mathrm{R, D, U}, H_2O_2)$ via SR, which are the reaction rate equations sought in this study. Compared with the settings for seeking the concentration-time relationships (Table 1), the settings for the SR of the reaction rate equations (f_{R} and f_{D}) apply a deeper search, which raises the maximum number of iterations from 100 to 150 and the population size from 100 to 200. Furthermore, the division operator and the exponential operators are disabled, as they are rarely seen in rate laws. Moreover, the complexity cap is tightened from 30 to 20, reflecting the parsimony of natural kinetics.

Figure 3 presents the resulting f_{R} and f_{D} candidates given by SR. Some mathematical expressions do not contain H_2O_2 in the terms, violating the knowledge that the reaction cannot start without its existence. Thus, the expressions without H_2O_2 are discarded.

Thereafter, an exhaustive mapping is carried out to construct all possible $(f_{\mathrm{R}}, f_{\mathrm{D}})$ combinations from expressions that pass the screening and derive the corresponding f_{U}. The resulting model candidates, consisting ordinary differential equations, are integrated with respect to time, giving the predicted $\hat{\mathrm{R}}, \hat{\mathrm{D}}$ and $\hat{\mathrm{U}}$ at the sampling times. Afterwards, the AIC values are calculated by leveraging the measured data of $\mathrm{R, D}$ and U. This procedure is graphically presented in Figure 4.

3.3. SIO Analysis & Best Model Selection

Utilizing the MATLAB-based tool STRIKE-GOLDD, the SIO of the 10 model candidates with the lowest AIC values are analyzed, where the models with structural defects will be reparametrized by AutoRepar to fix this issue. Note that the AIC of the reparametrized models are re-computed, since the number of model parameters have changed. Eventually, the best kinetic model is selected among these SIO-assured models based on the updated AIC values. The rate equations of the best model are given by Equations 7 to 10, with the parameter values presented in Table 2. The SIO of this model is guaranteed, as evidenced by the SIO analysis records presented in Figure 5. Moreover, this model has an outstanding prediction performance as demonstrated in Figure 6. Hence, a high-fidelity and structurally rigorous reaction rate model is obtained in this case study.

$$\dot{\mathrm{R}} = p_1 + (p_2 - p_3 \cdot \mathrm{R} \cdot \mathrm{D} \cdot H_2O_2) \cdot \mathrm{R} \tag{7}$$

$$\dot{\mathrm{D}} = (p_4 + p_5 \cdot \mathrm{R} \cdot \mathrm{D}) \cdot \mathrm{R} - p_6 \cdot H_2O_2 \tag{8}$$

$$\dot{\mathrm{U}} = -(p_1 + (p_2 - p_3 \cdot \mathrm{R} \cdot \mathrm{D} \cdot H_2O_2) \cdot \mathrm{R}) - ((p_4 + p_5 \cdot \mathrm{R} \cdot \mathrm{D}) \cdot \mathrm{R} - p_6 \cdot H_2O_2) \tag{9}$$

$$\dot{H_2O_2} = 2(p_1 + (p_2 - p_3 \cdot \mathrm{R} \cdot \mathrm{D} \cdot H_2O_2) \cdot \mathrm{R}) + ((p_4 + p_5 \cdot \mathrm{R} \cdot \mathrm{D}) \cdot \mathrm{R} - p_6 \cdot H_2O_2) \tag{10}$$

Table 2: Parameter estimates of the best rate model

Parameter	Value	Unit
p_1	0.194	$mol \cdot m^{-3} \cdot min^{-1}$
p_2	-5.79×10^{-2}	min^{-1}
p_3	-8.51×10^{-10}	$mol^{-3} \cdot m^{-9} \cdot min^{-1}$
p_4	6.01×10^{-2}	min^{-1}
p_5	8.72×10^{-7}	$mol^{-2} \cdot m^{-6} \cdot min^{-1}$
p_6	-3.62×10^{-3}	min^{-1}

4. CONCLUSION

An integrated workflow coupling SR with joint SIO analysis is proposed and demonstrated through a partially observed sulfide oxidation case study. By embedding scarce prior knowledge and limited experimental

```
--------------------------------
>>> STRIKE-GOLDD toolbox 3.0
--------------------------------

Analyzing the TempCombo_k0_0_162105 model...

>>> The model contains:
4 states:
[R; D; U; H2O2]
3 outputs:
[R; D; U]
0 known inputs:

0 unknown inputs:

6 parameters:
[p1; p2; p3; p4; p5; p6]

>>> Building the observability-identifiability matrix requires at least 3 Lie derivatives
    Calculating derivatives: 1 2 3
>>> Observability-Identifiability matrix built with 3 Lie derivatives
    (calculated in 2.425598e-01 seconds)
>>> Calculating rank...
    Rank = 9 (calculated in 6.448850e-02 seconds)
>>> Observability-Identifiability matrix built with 4 Lie derivatives
    (calculated in 3.031976e-01 seconds)
>>> Calculating rank...
    Rank = 10 (calculated in 1.766829e-01 seconds)

------------------------
>>> RESULTS SUMMARY:
------------------------

>>> The model is Fully Input-State-Parameter Observable (FISPO):
    All its states are observable.
    All its parameters are locally structurally identifiable.
Total execution time: 1.485672e+00


----------------------------------------
The model is already FISPO.
 There is no need for reparameterization.
----------------------------------------
```

Figure 5. SIO analysis record of the best model given by STRIKE-GOLDD.

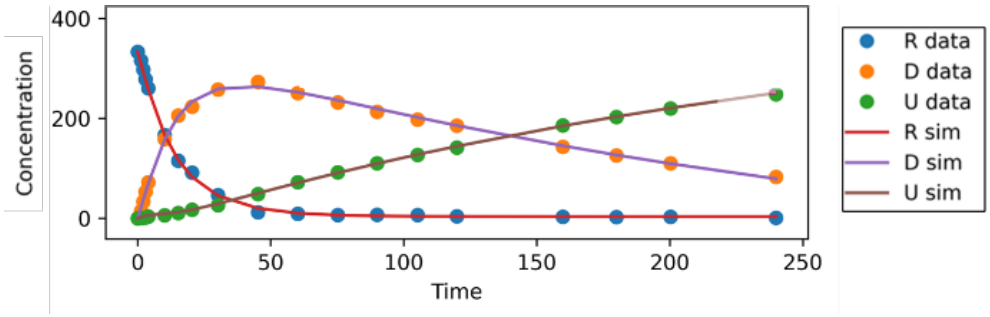

Figure 6. Fitting performance of the best model.

data, the approach returns a model that simultaneously fits the data and preserves rigorous structure. In contrast to purely data-driven SR pipelines, which often yield non-identifiable or structurally inconsistent models, the proposed framework systematically filters candidates based on identifiability and observability, leading to a substantially reduced and more interpretable model space. Relative to conventional mechanistic model

development, the workflow achieves comparable structural guarantees while requiring markedly less a priori specification, thereby improving robustness and credibility in early-stage process modelling where data and mechanistic insight are both limited.

ACKNOWLEDGEMENTS

The authors acknowledge funding from the UK Engineering and Physical Sciences Research Council (EPSRC), for Made Smarter Innovation – Digital Medicines Manufacturing Research Centre (DM2), EP/V062077/1.

AUTHOR IDENTIFIERS

Author ORCIDs:
Yuan X: 0009-0006-2211-4879
Benyahia B: 0000-0002-5397-1498

REFERENCES

1. Ramin, P, Mansouri, SS, Udugama, IA, Benyahia, B, Gernaey, KV (2018) Modelling continuous pharmaceutical and bio-based processes at plant-wide level: A roadmap towards efficient decision-making, *Chemistry Today*, 36(2), 26-30.
2. Ma Y, Yuan X, Yewale A, Benyahia B. Recent advances in the digital transformation of crystallization process development and operation: synergy between model- and ai-driven strategies. Particuology 110:223-253 (2026). https://doi.org/10.1016/j.partic.2026.01.014
3. Campbell, TJS, Rielly, CD, Benyahia, B (2024) Sustainability and Quality-by-Digital Design of an Integrated End-to-End Continuous Pharmaceutical Process. *Computer Aided Chemical Engineering*, 53, 343-348
4. Benyahia, B. (2018) Applications of a plant-wide dynamic model of an integrated continuous pharmaceutical plant: Design of the recycle in the case of multiple impurities. In Singh, R and Yuan, Z (ed) *Process Systems Engineering for Pharmaceutical Manufacturing: From Product Design to Enterprise-Wide Decisions*, Elsevier, pp.141-157, DOI: 10.1016/B978-0-444-63963-9.00006-3.
5. Mustoe CL, Turner AJ, Urwin SJ, Houson I, Feilden H, Markl D, Al Qaraghuli MM, Chong MWS, Robertson M, Nordon A, Johnston BF, Brown CJ, Robertson J, Adjiman C, Batchelor H, Benyahia B, Bresciani M, Burcham CL, Cardona J, Cottini C, Dunn AS, Fradet D, Halbert GW, Henson M, Hidber P, Langston M, Lee YS, Li W, Mantanus J, McGinty J, Mehta B, Naz T, Ottoboni S, Prasad E, Quist PO, Reynolds GK, Rielly C, Rowland M, Schlindwein W, Schroeder SLM, Sefcik J, Settanni E, Siddique H, Smith K, Smith R, Srai JS, Thorat AA, Vassileiou A, Florence AJ. Quality by digital design to accelerate sustainable medicines development. International Journal of Pharmaceutics 681:125625 (2025). https://doi.org/10.1016/j.ijpharm.2025.125625
6. Campbell, TJS, Rielly, CD, Benyahia, B (2024) Sustainability and Quality-by-Digital Design of an Integrated End-to-End Continuous Pharmaceutical Process. *Computer Aided Chemical Engineering*, 53, 343-348
7. Gao S, Benyahia B. Robust techno-economic analysis, life cycle assessment, and quality and sustainability by digital design of three alternative continuous pharmaceutical tablet manufacturing processes. Systems and Control Transactions 4:2568-2573 (2025). https://doi.org/10.69997/sct.104102
8. Liu, J., Benyahia, B. Systematic model-based dynamic optimization of a combined cooling and antisolvent multistage continuous crystallization process. Computer Aided Chemical Engineering 50, 1221-1227.
9. Benyahia B, Anandan PD, Rielly C. Robust model-based reinforcement learning control of a batch crystallization process. 2021 9th International Conference on Systems and Control (ICSC) :89-94 (2021). https://doi.org/10.1109/icsc50472.2021.9666494
10. Alam MN, Bhat S, Kodamana H, Rathore AS. A data-driven symbolic regression framework for modeling and multiobjective optimization of a microbial fermentation system. Ind. Eng. Chem. Res. 64:23220-23233 (2025). https://doi.org/10.1021/acs.iecr.5c02567
11. de Carvalho Servia MÁ, Sandoval IO, Hii KK(, Hellgardt K, Zhang D, Antonio del Rio Chanona E. The automated discovery of kinetic rate models – methodological frameworks. Digital Discovery 3:954-968 (2024). https://doi.org/10.1039/d3dd00212h
12. Doherty S, Knight JG, Carroll MA, Clemmet AR, Ellison JR, Backhouse T, Holmes N, Thompson LA, Bourne RA. Efficient and selective oxidation of sulfides in batch and continuous flow using styrene-based polymer immobilised ionic liquid phase supported peroxotungstates. RSC Adv. 6:73118-73131 (2016). https://doi.org/10.1039/c6ra11157b
13. Javadi AB, Pong P. A review on symbolic regression in power systems: methods, applications, and future directions. Renewable and Sustainable Energy Reviews 224:116075 (2025). https://doi.org/10.1016/j.rser.2025.116075
14. Koza J. Genetic programming as a means for programming computers by natural selection. Stat Comput 4: (1994).

https://doi.org/10.1007/bf00175355
15. Lima FARD, de Moraes MGF, Rebello CM, Barreto AG Jr, Secchi AR, de Souza MB Jr, Nogueira IBR. Interpretable and uncertainty-aware machine learning for trustworthy prediction in batch crystallization. Chemical Engineering and Processing - Process Intensification 215:110350 (2025). https://doi.org/10.1016/j.cep.2025.110350
16. Massonis G, Banga JR, Villaverde AF. Autorepar: a method to obtain identifiable and observable reparameterizations of dynamic models with mechanistic insights. Intl J Robust & Nonlinear 33:5039-5057 (2021). https://doi.org/10.1002/rnc.5887
17. Yuan X, Nagy ZK, Benyahia B. A comprehensive framework for model evaluation and refinement using mbdoe estimability and structural identifiability: application to a crystallization process. Chemical Engineering Journal Advances 23:100823 (2025). https://doi.org/10.1016/j.ceja.2025.100823