Integrating Knowledge-Guided Symbolic Regression for Model-Based Design of Experiments to Automate Process Flow Diagram Development

Alexander W. Rogers,a Amanda Lane,b Philip Martin,a Dongda Zhanga\*

aDepartment of Chemical Engineering, The University of Manchester, Oxford Road, Manchester, M1 3AL, UK

bUnilever R&D Port Sunlight, Quarry Road East, Bebington, CH63 3JW, UK

\*Corresponding author email:dongda.zhang@manchester.ac.uk

Abstract

New products must be formulated rapidly to succeed in the global formulated product market; however, key product indicators (KPIs) can be complex, poorly understood functions of the chemical composition and processing history. Consequently, process scale-up must undergo expensive trial-and-error campaigns that do not guarantee optimality. To accelerate process flow diagram (PFD) optimisation and knowledge discovery, this work proposed a novel digital framework to automatically quantify process mechanisms by integrating symbolic regression (SR) within model-based design of experiments (MBDoE). For each iteration, SR proposed a Pareto front of interpretable mechanistic expressions, and then MBDoE designed a new experiment to discriminate between them while automatically balancing the objective of PFD optimisation. To investigate the framework’s performance, a new process model capable of simulating general formulated product synthesis was constructed to generate in-silico data for different case studies. The framework could effectively discover ground-truth process mechanisms within a few iterations, indicating its great potential within the general chemical industry for digital manufacturing and product innovation.

**Keywords**: knowledge discovery, symbolic regression, model-based design of experiments, interpretable machine learning, process flow diagram optimisation.

* 1. Introduction

At present, the scale-up of formulated products must undergo expensive trial-and-error campaigns due to the complex, poorly understood link between the final product properties, chemical composition, and processing conditions during manufacture. At this moment, MBDoE is the most promising approach to solving this challenge, whereby a model is used to efficiently guide exploration vs. exploitation of the experimental design space. The general MBDoE framework is flexible. The model used can be a mechanistic, machine learning or hybrid model. Experiments can be designed to yield the most new statistical information for the minimum amount of time and resources or, if formulated as a multi-objective optimisation problem, be designed to simultaneously optimise operating conditions (Franceschini & Macchietto, 2008). However, using MBDoE for PFD development within the formulation and speciality industries remains a severe challenge due to insufficient high-quality data for pure machine learning methods or quantitative descriptions of the complex formulation processes for building hybrid or pure mechanistic models. As such, the best solution is to propose a general framework for automatically discovering good mechanistic models – an approach that would be interpretable. By their construction, analytical expressions can be inspected, debugged, and adapted by expert practitioners to incorporate prior physical knowledge to improve data efficiency or discover new physical knowledge.

The sparse identification of nonlinear dynamics (SINDy) algorithm (Brunton et al., 2016) promotes sparsity among a library of candidate functions to discover ordinary differential equations (ODEs). However, the dynamics must have a sparse representation in the pre-defined library. Genetic algorithms for symbolic regression (SR) explore a much larger space of expressions by selection, mutation, and crossover defined only by a set of input features and mathematical operators (de Franca et al., 2023); as such, SR has helped discover constitutive property relationships (Angelis et al., 2023) and has been applied to discovering kinetic rate models for catalytic processes (Servia et al., 2023). However, without prior knowledge to constrain the solution space, it is very challenging for SR to find accurate expressions for complex systems – even then, the identified expression may represent a local approximation, reducing its physical interpretability. Hence, there have been some, albeit very few, attempts to incorporate prior physical knowledge into SR (Kronberger et al., 2022; Reinbold et al., 2021), so this remains an open challenge. Therefore, this work proposes a novel digital modelling framework integrating SR within MBDoE to aid automatic knowledge discovery and PFD optimisation. This framework is designed to efficiently recover underlying governing equations representing the scale-independent process dynamics through an iterative procedure. To help accelerate system identification and minimise the number of experiments required, the structure of the expressions searched by SR is also constrained based on prior physical knowledge.

* 1. Methodology

The framework integrating knowledge-guided SR and MBDoE is illustrated in Fig. (1). Starting at Step 1, an initial set of experiments is conducted based on expert experience of important PFD parameters, $ϑ$ , and their lower, $ϑ\_{lb}$, and upper, $ϑ\_{ub}$, bounds.



Figure 1: General SR-MBDoE for proposing mechanistic expressions for the underlying system and designing experiments that automatically balance exploration vs. exploitation.

In Step 2, SR identifies a Pareto set of expressions balancing fitting accuracy and complexity for the intrinsic dynamics, $dψ/dt=f\left(X,P\right)$. Where $ψ$, $X$ and $P$ are vectors of KPIs, chemical concentrations and processing conditions, respectively. In Step 4, MBDoE will design a new experiment to discriminate between the candidate expressions by minimising the multi-objective function, $J\left(ϑ\right)$, which can prioritise knowledge discovery alone or simultaneously explore opportunities for PFD optimisation. These objectives are re-weighted automatically in Step 3 based on the information and optimality gains from the experiment in the prior MBDoE iteration.

* + 1. Knowledge-Guided Symbolic Regression

In Step 2, tournament selection promotes and mutates the candidates with the smallest $L\_{MSE}^{'}$ from a population of expressions using the Python-Julia library PySR (Cranmer, 2023). In the end, the fittest individuals in the population at each level of complexity were lined up as a Pareto set and scored by the negated derivative of the log-loss with respect to complexity, as in Eq. (1) (Cranmer, 2023). Where $L\_{MSE}$ is the mean-square fitting error (MSE) for each expression, while the superscript $C$ is the complexity of the expression. The top scoring expressions represent a balance between fitting accuracy and complexity.

$$\begin{array}{c}Score=-Δlog\left(L\_{MSE}^{C}\right)/ΔC\#\left(1\right)\end{array}$$

However, model selection by parsimony alone does not guarantee that the selected model will extrapolate well, which is key to minimising the number of experiments for system identification; for this, prior knowledge about the nature of the function is necessary.

$$\begin{array}{c}L\_{MSE}^{'}=L\_{MSE}+P\#\left(2a\right)\end{array}$$

$$\begin{array}{c}P=\left\{\begin{matrix}\infty ,&if G\notin G^{'}=k\left(X,P\right)×\left[f\left(X,P\right)-b\left(X,P\right)÷K\left(X,P\right)\right]\\0,&otherwise\end{matrix}\right.\#\left(2b\right)\end{array}$$

Expressions were guided towards structures with physically meaningful interpretations. Appended to $L\_{MSE}$ used to evaluate fitness during tournament selection, was a penalty, $P$, defined in Eq. (2b) to take infinity when the expression, $G$, was not of the form, $G^{'}$. Generalisable to PFD optimisation is the assumption of state equilibration, where $k\left(⋅\right)$, $f\left(⋅\right)$, $b\left(⋅\right)$ and $K\left(⋅\right)$ are functions representing the overall rate, forward and backward driving forces, and equilibrium constant for the underlying formulation process, but $G^{'}$can take other structural forms if different prior knowledge is considered true.

* + 1. Multi-Objective MBDoE for PFD Development

In Step 4, MBDoE designs new experiments by minimising the multi-objective function, $J\left(ϑ\right)$, in Eq. (3). $J\_{M}\left(ϑ\right)$ is the expected information gain, $J\_{P}\left(ϑ\right)$ is the objective function for process optimisation, while $J\_{M}^{max}$ and $J\_{P}^{max}$ are normalisation constants. $0 \leq α\leq 1$ systematically balances these objectives and is updated in Step 3 each MBDoE iteration.

$$\begin{array}{c}\min\_{ϑ}J\left(ϑ\right)=α⋅\frac{J\_{M}\left(ϑ\right)}{J\_{M}^{max}}+\left(1-α\right)⋅\frac{J\_{P}\left(ϑ\right)}{J\_{P}^{max}}\#\left(3a\right)\end{array}$$

$$\begin{array}{c}s.t. ϑ\_{lb}\leq ϑ\leq ϑ\_{ub}\#\left(3b\right)\end{array}$$

$J\_{M}\left(ϑ\right)$ was estimated as the variance in the final product KPI, $\hat{ψ}\_{f}$, predicted by simulating the PFD with combinations of the top three scoring kinetic expressions proposed by SR selected by Eq. (1). To balance exploration vs. exploitation automatically, in Step 3, $α$ is re-weighted using Eq. (4) based on the actual information gained, $ΔJ\_{M}$, by the experiment and how far from optimality, $ΔJ\_{P}$, was the PFD in the previous MBDoE iteration.

$$\begin{array}{c}α=\frac{ΔJ\_{M}}{ΔJ\_{M}+ΔJ\_{P}}\#\left(4\right)\end{array}$$

* 1. Formulation Process Modelling and Case Study

A liquid product, typical of cosmetic and pharmaceutical creams, was used as a case study, and a new mechanistic model was proposed for the first time to approximate its formulation and KPI dynamics. This model was used to run computational experiments and generate in-silico data to test the SR-MBDoE framework.

$$\begin{array}{c}r\_{1}= k\_{1}⋅\dot{γ}⋅\left(α-T\right)⋅\left[X\_{A}X\_{W}\right]⋅H\left(T-T\_{K}\right)\#\left(5a\right)\end{array}$$

$$\begin{array}{c}r\_{2}=k\_{2}⋅\dot{γ}⋅T⋅\left[X\_{L}X\_{W}-\frac{X\_{L^{\*}}}{K\_{2}⋅T^{-1}}\right]\#\left(5b\right)\end{array}$$

$$\begin{array}{c}r\_{3}=k\_{3}⋅\dot{γ}⋅\left[ X\_{L} -\frac{X\_{V}}{K\_{3}⋅\dot{γ}⋅\left(T-β\right)}\right]⋅H\left(T-T\_{K}\right)\#\left(5c\right)\end{array}$$

The model grouped the chemical constituents into five phases (i.e., $W$, $A$, $L$,$L^{\*}$ and $V$); proposed rate equations for three mechanisms: $r\_{1}$: $2A+5W\rightarrow L$ and $r\_{2}: L+10W⇌L^{\*}$ and $r\_{3}: 3L⇌V$; and embedded the kinetics into the recycle emulsification configuration in Fig. (2a). The rates, $r\_{i}$, in Eq. (5) were functions of $X$ and $P=\left[T,\dot{γ}\right]$, where $T$ was temperature and $\dot{γ}$ was average shear rate, while $H\left(T\right)$ was the Heaviside switch function and $k\_{i}$, $K\_{i}$, $α$, $β$ and $T\_{K}$ were constants. At any time, KPI was: $ψ=f\left(X\right)$. Fig. (2b) shows the in-silico product KPI profiles for the four experiments that initiated MBDoE in the following case studies. This modelling approach is generally applicable to approximating product formulation by sequential ingredient additions and processing operations.



Figure 2: Recycle configuration for product manufacturing (a) and four example in-silico process dynamic KPI profiles (b), where $V$ is volume, $Q$ is volumetric flowrate and $T$ is temperature; the superscripts $w$, $a$, $s$, $p$ and $j$ denote two of the phases, the in-line mixer.

* 1. Results and Discussion

The SR-MBDoE framework was investigated in Case Study 1 when the sole focus was process knowledge discovery: $α=1$; then in Case Study 2 when the aim was simultaneous knowledge discovery and PFD optimisation: $α=f\left(ΔJ\_{M}, ΔJ\_{P}\right)$. In-silico experiments measured $∆X$and$P$over the in-line mixer each minute where the intrinsic dynamics dominated. For each MBDoE iteration, SR proposed three candidate expressions for each of the three rate equations $r\_{i}=f\left(X,P\right)$ (i.e., nine overall); then, a new experiment was designed and conducted by minimising $J\left(ϑ\right)$, expanding the dataset.

* + 1. Case Study 1: Process Knowledge Discovery

For each MBDoE iteration ($I\_{MBDoE}$), Case Study 1A proposed a new set of equations from scratch, while Case Study 1B carried over and improved the equations from the previous $I\_{MBDoE}$, representing a trade-off between the risk of inherited biases and computational efficiency. Fig. (3a) and (3b) show the prediction MAPE for the product KPI and fitting MSE for the top three scoring expressions for each rate equation following each $I\_{MBDoE}$ for Case Studies 1A and 1B, respectively. To begin with, the “evolutionary pressure” towards the ground truth was weak, and there were too many similarly fitting expressions to search. As more carefully designed experiments were added, the difference in the MSE between correct and incorrect expressions during tournament selection became stronger, encouraging the promotion of better-fitting expressions.



Figure 3: Mean absolute percentage error (MAPE, for prediction) and mean-square error (MSE, for fitting) of the expressions $r\_{i}^{j}$ proposed at each iteration of Case Studies 1A (a) and 1B (b); $i$ and $j$ index the rate equation and relative complexity. Bars corresponding to the top-scoring expressions used for estimating the MAPE at each iteration are hashed.

In both Case Studies, the MAPE peaked. In Case Study 1B, the ground truth for $r\_{3}$ was found in $I\_{MBDoE}=4$ but then demoted but not forgotten in $I\_{MBDoE}=5$ in favour of a worse-fitting approximation with an MSE of 0.0086, but its much lower complexity gave it a higher score of 1.3 compared with 1.1 and 0.82 for the other candidates with MSEs of 0.00001 and 0.0009, respectively. The choice of metric to score and select expressions is critical; however, each metric or information criterion (e.g., Akaike, Bayesian and Hannan-Quinn) comes with its own biases. Only carefully designed experiments can reliably discriminate between different hypotheses and identify the correct model structure. Thus, by $I\_{MBDoE}=6$ when the new experiment pushed up the MSE of the approximation from 0.0086 to 0.011, the score ranking flipped, and the ground truth was re-identified. By contrast, the MAPE never recovered in Case Study 1A, suggesting that building complete expressions from scratch is more challenging. So, rather than inherited biases hindering discovery, adding new information by MBDoE incrementally to improve existing expressions can guide SR towards the correct structure more efficiently.

With the constraint (i.e., based on prior knowledge of state equilibration) active, SR successfully yielded physically insightful expressions that could be interpreted in terms of key forward and backward driving force factors. Now, Case Study 1C tested the performance of the SR-MBDoE framework when lifting this constraint. Here, the MAPE decreased from 25.6% to 1.98% and plateaued for $I\_{MBDoE}\geq 7$, demonstrating that constraining the search to the correct structure significantly improved fitting accuracy and sped up knowledge discovery when the number of experiments was small.

* + 1. Case Study 2: Simultaneous Knowledge Discovery and PFD Optimisation

The aim was to hit a target final product KPI, $ψ\_{t}$, within a $\pm 3\%$ tolerance of a specific value while minimising total batch time, $τ$. Initially, MBDoE bounced between exploration ($α>0.5$) and exploitation ($α<0.5$). For as long as $ΔJ\_{M}$ was large, and new experiments continued to be informative, then exploration was prioritised. Once $ΔJ\_{M}$ became small ($α=0.2$ by $I\_{MBDoE}=5$), and new experiments no longer proved to be as informative, then process optimisation was prioritised. If, at any point, the model aimed for and successfully hit on a good recipe (i.e., one that achieved an in-spec KPI), then there would be nothing new to learn or improve about the process within the local vicinity; thus, the next iteration would bounce back to pure exploration. By $I\_{MBDoE}=7$ the nominal total batch time was greatly reduced from $τ=100 min$ at $I\_{MBDoE}=0$ to $τ=57 min$, while the final KPI satisfied the requirement, further evidencing the practical advantage and efficiency of the SR-MBDoE framework. By discovering mechanistic rate expressions, the physical trade-offs made by the optimised PFD were also interpretable.

* 1. Conclusions

Through two case studies, despite the highly complex nature of the underlying ground truth, the proposed knowledge-guided SR-MBDoE framework could recover the ground truth exactly after only a small number of experiments, demonstrating its great potential. While carrying over expressions from previous MBDoE iterations for modification proved more successful than building expressions from scratch. However, selecting expressions based on statistical parsimony alone risks bias; only carefully designed experiments can reliably discriminate between similarly fitting candidates of different complexities. Then, when the knowledge-guided constraint on the expressions’ structure was lifted, the prediction accuracy for the same number of experiments decreased substantially. By synergising human intelligence with the automatic discovery and discrimination of interpretable mechanistic models representing the scale-independent process dynamics, the proposed framework shows excellent potential for accelerating product innovation, scale-up and design of PFDs for producing new formulations.

References

Angelis, D., Sofos, F., & Karakasidis, T. E. (2023). Artificial Intelligence in Physical Sciences: Symbolic Regression Trends and Perspectives. *Archives of Computational Methods in Engineering*, *30*(6), 3845–3865. https://doi.org/10.1007/s11831-023-09922-z

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *113*(15), 3932–3937. https://doi.org/10.1073/pnas.1517384113

Cranmer, M. (2023). *Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl*. https://doi.org/10.48550/ARXIV.2305.01582

de Franca, F. O., Virgolin, M., Kommenda, M., Majumder, M. S., Cranmer, M., Espada, G., Ingelse, L., Fonseca, A., Landajuela, M., Petersen, B., Glatt, R., Mundhenk, N., Lee, C. S., Hochhalter, J. D., Randall, D. L., Kamienny, P., Zhang, H., Dick, G., Simon, A., … La Cava, W. G. (2023). *Interpretable Symbolic Regression for Data Science: Analysis of the 2022 Competition* (arXiv:2304.01117). arXiv. http://arxiv.org/abs/2304.01117

Franceschini, G., & Macchietto, S. (2008). Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science*, *63*(19), 4846–4872.

Servia, M. Á. de C., Sandoval, I. O., Hellgardt, K., Kuok, K., Hii, Zhang, D., & Chanona, E. A. del R. (2023). *The Automated Discovery of Kinetic Rate Models—Methodological Frameworks* (arXiv:2301.11356). arXiv. http://arxiv.org/abs/2301.11356