12th IFAC International Symposium on
Advanced Control of Chemical Processes
July 14-17, 2024. Toronto, Canada

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

Alexander W. Rogers*. Amanda Lane. Cesar Mendoza**. Simon Watson****

Adam Kowalski. Philip Martin*. Dongda Zhang***

**Department of Chemical Engineering, University of Manchester, Oxford Road, Manchester, M1 3AL, UK (e-mail: dongda.zhang@manchester.ac.uk).*

***Unilever R&D Port Sunlight, Quarry Road East, Bebington, CH63 3JW, 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 currently undergo expensive trial-and-error campaigns. To accelerate process flow diagram (PFD) optimization and knowledge discovery, this work proposes a novel digital framework to automatically quantify process mechanisms by integrating symbolic regression (SR) within model-based design of experiments (MBDoE). Each iteration, SR proposed a Pareto front of interpretable mechanistic expressions, and then MBDoE designs a new experiment to discriminate between them while automatically balancing the objective of PFD optimization. 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 optimization.

1. INTRODUCTION

The global formulated products industry is large but competitive and dynamic, requiring rapid development of new products. However, the final product properties are often complex, poorly understood functions of the chemical composition and the history of processing conditions during manufacture. Hence, new product development and scale-up must undergo expensive trial-and-error campaigns that do not guarantee economic or environmental process optimality.

At this moment, model-based design of experiments (MBDoE) is the most promising approach to solving this challenge, whereby a model is used to guide exploration vs. exploitation of the experimental design space efficiently. 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 (Franceschini and Macchietto, 2008). If formulated as a multi-objective optimization problem, experiments can also be designed to discover new knowledge and optimize operating conditions simultaneously (Echtermeyer *et al.*, 2017).

However, using MBDoE for process flow diagram (PFD) development within the formulation and specialty 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.

In recent years, there has been a push towards parsimonious analytical expressions with the lowest complexity required to describe the main features of the data to avoid overfitting. The sparse identification of nonlinear dynamics (SINDy) algorithm (Brunton, Proctor and Kutz, 2016) promotes sparsity among a library of candidate functions to discover ordinary differential equations (ODEs). However, such algorithms rely on the dynamics having a sparse representation in a pre-defined library. This has motivated genetic algorithms for symbolic regression (SR) that 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 had success in discovering constitutive property relationships (Angelis, Sofos and Karakasidis, 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 simply represent a local approximation, reducing its physical interpretability. Hence, there have been some, albeit very few, attempts to incorporate prior physical knowledge

Copyright © 2024 the authors. Accepted by IFAC for publication under a Creative Commons License CC-BY-NC-ND.

into SR (Kronberger et al., 2022; Reinbold et al., 2021), so this topic remains an open challenge.

Therefore, this work proposes a novel digital modelling framework integrating SR within MBDoE to aid automatic knowledge discovery and process flow diagram (PFD) optimization. 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 minimize the number of experiments required, the structure of the expressions searched by SR is constrained based on prior physical knowledge.

2. METHODOLOGY

The general framework integrating knowledge-guided SR and MBDoE is illustrated in Fig. 1 as a flowchart.

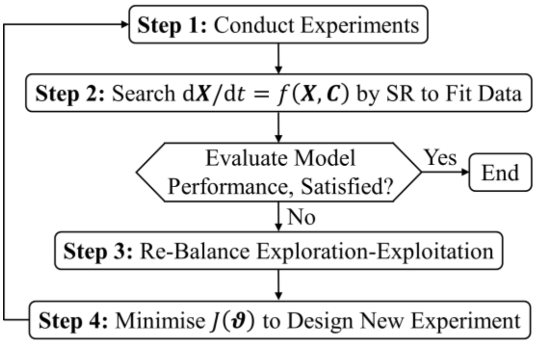

Step 1: Conduct Experiments
Step 2: Search $\mathrm{d}\boldsymbol{X}/\mathrm{d}t = f(\boldsymbol{X}, \boldsymbol{C})$ by SR to Fit Data
Evaluate Model Performance, Satisfied?
Yes
End
No
Step 3: Re-Balance Exploration-Exploitation
Step 4: Minimise $J(\boldsymbol{\vartheta})$ to Design New Experiment

Figure 1. General SR-MBDoE flowchart for proposing expressions designing new experiments that balance exploration vs. exploitation.

Upon starting at Step 1, an initial set of experiments is conducted based on experience or understanding of important PFD parameters, $\boldsymbol{\vartheta}$, and their bounds, boxing the experimental design space. In Step 2, SR identifies a Pareto set of expressions balancing fitting accuracy and complexity for the dynamics, $\mathrm{d}\boldsymbol{X}/\mathrm{d}t = f(\boldsymbol{X}, \boldsymbol{C})$. Where $\boldsymbol{X}$ and $\boldsymbol{C}$ are vectors of chemical concentrations (e.g., in this work these are grouped phase concentrations for a multi-phase formulated product) and processing conditions (e.g., shear rate and temperature). Also observed is the KPI that is a function: $\boldsymbol{\varphi} = h(\boldsymbol{X})$. Expressions are scored and selected to build process models. If the top scoring model or PFD performance is satisfactory, then MBDoE should be terminated, otherwise, a new experiment is conducted. In Step 4, experiments are designed by minimizing the objective function, $J(\boldsymbol{\vartheta})$. The objective function, $J(\boldsymbol{\vartheta})$, can design experiments for maximum model discrimination and therefore knowledge discovery, process optimization, or balance both simultaneously. The weighting between these two objectives is automatically controlled in Step 3 based on the information and optimality improvement in the experiment from the previous MBDoE iteration.

2.1 Knowledge-Guided Symbolic Regression

With reference to Fig. 1, Step 2 is now detailed.

2.1.1. Symbolic Regression

In this work, tournament selection promoted and mutated the best candidates from a population of expressions represented by directed acyclic graphs using the Python-Julia library PySR by (Cranmer, 2023). PySR embeds the genetic algorithm inside an *evolve-simplify-optimize* loop: after a set number of tournaments and mutations, the equations are simplified using algebraic equivalencies, followed by a few iterations of local-gradient-based optimization to refine the numerical constants in the expressions. In the end, the fittest individuals in the population at each level of complexity are lined up as a Pareto set and scored by the negated derivative of the log-loss with respect to complexity, as shown in (1b) (Cranmer, 2023). In (1a) $\mathcal{L}_i$ is the MSE between the predicted $\hat{\boldsymbol{y}}_i \in \mathbb{R}^{N\times 1}$ and measured $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ outputs weighted by the diagonal matrix $\boldsymbol{\Lambda}$ averaged over N datapoints, where the subscript i indexes the candidate expression from the Pareto set of expressions. Complexity, $\mathcal{C}$, is defined as the total number of operators, variables, and constants. For the ordered Pareto set: $\mathcal{C}_{i+1} > \mathcal{C}_i$.

$$\mathcal{L}_i = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{y} - \hat{\boldsymbol{y}}_i)^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{y} - \hat{\boldsymbol{y}}_i) + P_i \qquad (1a)$$

$$\text{Score} = -\frac{\log(\mathcal{L}_{i+1}) - \log(\mathcal{L}_i)}{\mathcal{C}_{i+1} - \mathcal{C}_i} \qquad (1b)$$

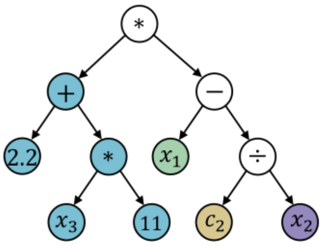
2.1.2. Constrained Structure

Good extrapolation is key to minimizing the number of experiments needed for system identification. However, without correct underlying theory with which to motivate model selection or construction, this is not guaranteed. There can be multiple possible expressions that fit the observed data well but then disagree further away, without necessarily being accurate. Information criteria (e.g., Akaike, Bayesian and Hannan-Quinn) have been proposed for selecting models based on the likelihood of observing data given a certain model while penalizing the number of parameters. However, these do not guarantee correct model selection. Consider fitting data points over a narrow slice of an input domain; without prior knowledge about the nature of the underlying function, a linear correlation would indeed be the simplest, well-fitting hypothesis. For this reason, physical models motivated from correct underlying theory tend to extrapolate better than statistical models that are not motivated by physical theory.

$$P_i = \begin{cases} \infty, & \text{if } G_i \notin k(\cdot) \times [f(\cdot) - b(\cdot) \div K(\cdot)] \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

Therefore, expressions were constrained to those that adhere to some physical interpretation, discarding nonconformers. Implemented by means of the penalty, P, appended to (1a) defined in (2) takes infinity when the expression tree, G, is not in the set of desired structures. The penalty was implemented as a logical condition that examines the top-level operators (e.g., if the first operator is not multiplication: return infinity, otherwise return zero). Specific to the current case, G was constrained to take the form shown in Fig. 2, reflecting the general form of the kinetic equations in (8a), (8b) and (8c), constructed from forward and backward contributions to each mechanism. $k(\cdot)$, $f(\cdot)$, $b(\cdot)$ and $K(\cdot)$ are sub-expressions representing the overall rate constants, forward and backward driving forces, and equilibrium constants, respectively, for each of the underlying formulation process mechanisms.

These sub-expressions were found simultaneously, made partially identifiable by a cap on total complexity. Through this knowledge-guided evolution approach, one can expect SR to be more likely to discover physically insightful expressions. Note that while the assumption of state equilibration is generally applicable to formulation process PFD development, as explored here, the expression tree can be restricted to other structural forms if different prior knowledge is considered true.



$$k(\boldsymbol{X},\boldsymbol{C}) * \left[f(\boldsymbol{X},\boldsymbol{C}) - \frac{b(\boldsymbol{X},\boldsymbol{C})}{K(\boldsymbol{X},\boldsymbol{C})}\right]$$

Figure 2: Example tree of the form: $k(\cdot) \times [f(\cdot) - b(\cdot) \div K(\cdot)]$ where $\boldsymbol{X} = [x_1, x_2, x_3 \dots]^T$ and $\boldsymbol{C} = [c_1, c_2, c_3 \dots]^T$ are vectors of species concentrations and processing conditions, respectively.

2.2. *Model-Based Design of Experiments*

With reference to Fig. 1, Steps 3 and 4 are now detailed.

2.2.1. *General SR-MBDoE Algorithm*

Scientific models encode hypotheses; again, because multiple possible models will often fit the observed data well but then disagree further away, it is a fundamental part of the scientific method to design and conduct experiments to discriminate between these hypotheses. Following the flowchart in Fig. 1 from Step 1, an initial set of experiments is conducted based on experience or understanding of important PFD parameters, $\boldsymbol{\vartheta}$, (e.g., ingredient additions and processing conditions) and their upper, $\boldsymbol{\vartheta}_{\text{ub}}$, and lower, $\boldsymbol{\vartheta}_{\text{lb}}$, bounds boxing the experimental design space. Then, in Step 2, SR identifies different potential expressions for the intrinsic dynamics, $\mathrm{d}\boldsymbol{X}/\mathrm{d}t = f(\boldsymbol{X},\boldsymbol{C})$ balancing fitting accuracy and complexity. In Step 4, MBDoE designs new experiments by minimising the multi-objective function, $J(\boldsymbol{\vartheta})$, in (3). Where $J_E(\boldsymbol{\vartheta})$ is the exploration objective and $J_O(\boldsymbol{\vartheta})$ is the process optimization objective, while $J_E^{\max}$ and $J_O^{\max}$ are normalization constants. $0 \le \alpha \le 1$ systematically re-balances these two objectives and is updated in Step 3 for each MBDoE iteration.

$$\min_{\boldsymbol{\vartheta}} J(\boldsymbol{\vartheta}) = \alpha \cdot \frac{J_E(\boldsymbol{\vartheta})}{J_E^{\max}} + (1-\alpha) \cdot \frac{J_O(\boldsymbol{\vartheta})}{J_O^{\max}} \tag{3a}$$

$$J_E^{\max} = \max_{\boldsymbol{\vartheta}} J_E(\boldsymbol{\vartheta}) \tag{3b}$$

$$J_O^{\max} = \max_{\boldsymbol{\vartheta}} J_O(\boldsymbol{\vartheta}) \tag{3c}$$

$$\text{s.t. } \boldsymbol{\vartheta}_{lb} \le \boldsymbol{\vartheta} \le \boldsymbol{\vartheta}_{ub} \tag{3d}$$

2.2.2. *Information and Optimality Gain in PFD Development*

The exploration objective, J_E, is designed to maximize the information of new experiments for model discrimination. The PFD parameters, $\boldsymbol{\vartheta}$, prescribe the sequence of ingredient addition flowrates and processing conditions. Simulating the recipe from the start of the batch (i.e., $t = 0$) until the end of the batch (i.e., $t = \tau$) by integrating $\mathrm{d}\boldsymbol{X}/\mathrm{d}t = f(\boldsymbol{X},\boldsymbol{C})$ a matrix of predicted concentration profiles, $\widehat{\boldsymbol{X}}$, is obtained as in (4a). This is repeated using each set, S, of three top-scoring candidates for $\mathrm{d}\boldsymbol{X}/\mathrm{d}t = f(\boldsymbol{X},\boldsymbol{C})$ proposed by SR. Then the KPI, $\widehat{\boldsymbol{\psi}}_S$, is computed as a function of $\widehat{\boldsymbol{X}}_S$, as in (4b); this can either be known, or also constructed by SR. Finally, J_E is calculated as the variance in the predicted final product KPI, $\widehat{\boldsymbol{\psi}}_S$, as in (4c). The superscript k indexes different KPIs (e.g., physical properties) to be maximized for parallel experiments.

$$\widehat{\boldsymbol{X}}_S = \int_0^{\tau} \frac{\mathrm{d}\boldsymbol{X}}{\mathrm{d}t}(\boldsymbol{X},\boldsymbol{C})\bigg|_S \,\mathrm{d}t \tag{4a}$$

$$\widehat{\boldsymbol{\psi}}_S = h\big(\widehat{\boldsymbol{X}}_S\big) \tag{4b}$$

$$J_E^k = -\mathrm{var}\big(\big[\hat{\psi}_1^k, \hat{\psi}_2^k, \dots \hat{\psi}_{NS}^k\big]\big) \tag{4c}$$

For this case study, J_O, was defined in (5a) and (5b) to be the total batch time, τ, plus a quadratic penalty for when the KPI, $\hat{\psi}^k$, was not within tolerance, κ, of the target KPI, ψ_t^k.

$$J_O^k = \tau + \max(\varepsilon^k, 0) \tag{5a}$$

$$\varepsilon^k = \big(\psi_t^k - \hat{\psi}^k\big)^2 - (\kappa \cdot \psi_t^k)^2 \tag{5b}$$

Given the need to consider both objective functions, it is of critical importance to balance exploration vs. exploitation automatically; (6) systematically weights how much each objective was prioritized from one MBDoE iteration to the next as a function of ΔJ_M and ΔJ_P, based on the outcome of the experiments conducted in the previous MBDoE iteration.

$$\alpha = \frac{\Delta J_E}{\Delta J_E + \Delta J_O} \tag{6}$$

ΔJ_E in (7a) was the error between the predicted, $\hat{\psi}^k$, and actual, ψ_a^k, final KPI. ΔJ_O in (7b) and (7c) was the error between the target, ψ_t^k, and actual, ψ_a^k, final product KPI. Initially, $\alpha = 0.5$ for $I_{\text{MBDoE}} = 1$. As $\Delta J_E \ll \Delta J_O$, optimization for an on-spec product rather than exploration is increasingly preferred.

$$\Delta J_E = \big(\psi^k - \hat{\psi}_a^k\big)^2 \tag{7a}$$

$$\Delta J_O = \tau + \max(\varepsilon_k', 0) \tag{7b}$$

$$\varepsilon_k' = (\psi_t^k - \psi_a^k)^2 - (\kappa \cdot \psi_t^k)^2 \tag{7c}$$

3. FORMULATION PROCESS CASE STUDY

This research used liquid products, typical of cosmetic and pharmaceutical creams, as a case study, where no quantitative mechanistic model has yet been proposed to simulate these processes, presenting a severe challenge within the formulation industry for future digital manufacturing. To link the effect of chemical composition, manufacturing scale and different processing variables, a new mechanistic model was proposed for the first time to approximate the product formulation and KPI dynamics. This model was used to run computational experiments and generate in-silico data to test the SR-MBDoE methodology for knowledge discovery and simultaneous PFD optimisation.

The model grouped the chemical constituents into five phases. The rate, r_i, of material transformation due to mixing was described by (8a), (8b) and (8c) in terms of the concentrations of the five phases: $\boldsymbol{X} = [X_W, X_A, X_L, X_V, X_{L^*}]^T$ and processing conditions $\boldsymbol{C} = [T, \dot{\gamma}]^T$ representing the temperature, T, and the average shear rate, $\dot{\gamma}$, in the impeller region; the KPI was a function of composition, $\psi \propto X_{L^*}$; while k_i and K_i were the rate and equilibrium constants, respectively, and $H(T') = 0$ for $T' < 0$ but $H(T') = 1$ for $T' \geq 0$ where $T_K = 55$°C.

$$r_1 = k_1 \cdot \dot{\gamma} \cdot (\alpha - T) \cdot [X_A X_W] \cdot H(T - T_K) \tag{8a}$$

$$r_2 = k_2 \cdot \dot{\gamma} \cdot T \cdot \left[X_L X_W - \frac{X_{L^*}}{K_2 \cdot T^{-1}}\right] \tag{8b}$$

$$r_3 = k_3 \cdot \dot{\gamma} \cdot \left[X_L - \frac{X_V}{K_3 \cdot \dot{\gamma} \cdot (T - \beta)}\right] \cdot H(T - T_K) \tag{8c}$$

The components of $\mathrm{d}\boldsymbol{X}/\mathrm{d}t = f(\boldsymbol{X}, \boldsymbol{C})$ are defined in (9a), (9b), (9c), (9d) and (9e) in terms of the three reaction rates from (8a), (8b) and (8c), and the product stoichiometry. The intrinsic kinetics were embedded into a mass and energy balance for the recycle emulsification configuration depicted in Fig. 3a.

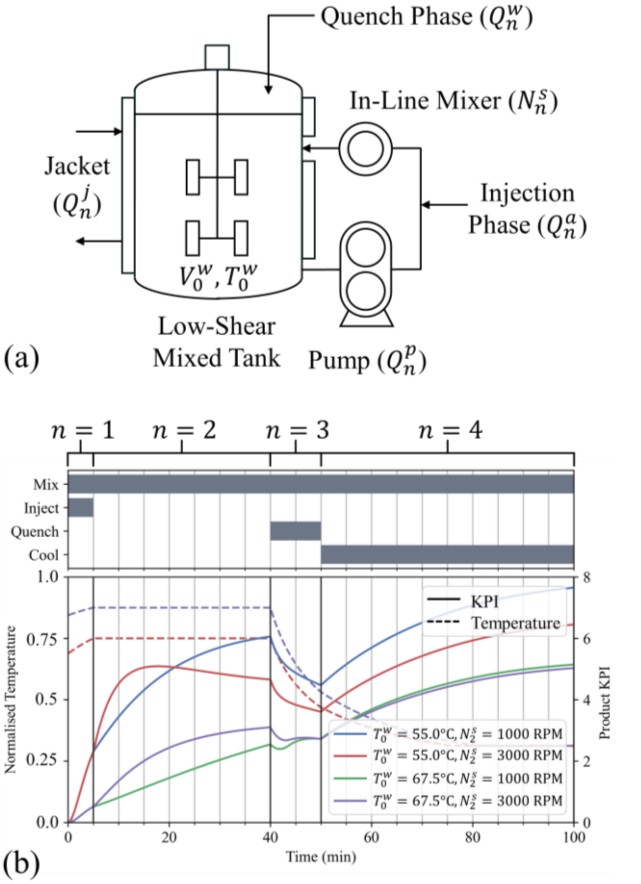


Figure 3. Equipment for formulated product manufacturing (a) where V is volume, Q is flowrate and T is temperature; w, a, s, p and j denote two ingredient streams, the in-line mixer, pump, and jacket, respectively. Shown are four simulated KPI profiles from following a four-step PFD (b) where n indexes the processing steps.

The synthetic PFD generated from the mechanistic model mirrors the four main steps involved in production. Fig. 3b shows the in-silico product KPI profiles for the four experiments used to initiate MBDoE in the following case studies. This approach of grouping chemical constituents into lumped phases to be described by a set of differential equations and mapping their concentrations to product KPIs is generally applicable to approximating formulated product synthesis by sequential ingredient additions and operations.

$$\frac{\mathrm{d}X_A}{\mathrm{d}t} = -2r_1 \tag{9a}$$

$$\frac{\mathrm{d}X_W}{\mathrm{d}t} = -5r_1 - 10r_2 \tag{9b}$$

$$\frac{\mathrm{d}X_L}{\mathrm{d}t} = +r_1 - r_2 - 3r_3 \tag{9c}$$

$$\frac{\mathrm{d}X_{L^*}}{\mathrm{d}t} = +r_2 \tag{9d}$$

$$\frac{\mathrm{d}X_V}{\mathrm{d}t} = +r_3 \tag{9e}$$

4. RESULTS AND DISCUSSION

The performance of the SR-MBDoE framework was investigated in Case Study 1 when the sole focus was knowledge discovery, then in Case Study 2 when the aim was simultaneous knowledge discovery and PFD optimisation.

4.1. Case Study 1: Process Knowledge Discovery

Each in-silico experiment took measurements representing the step change in the grouped phase concentrations, $\Delta\boldsymbol{X}$, over the in-line mixer, where the intrinsic dynamics dominate, and the local conditions experienced by the fluid, $\boldsymbol{C} = [T, \dot{\gamma}]^T$. If the residence time inside the in-line mixer is small, $\mathrm{d}\boldsymbol{X}/\mathrm{d}t$ can be approximated. Hence, the left-hand-side of (9a), (9b), (9c), (9d) and (9e) were known, and r_1, r_2 and r_3 could be found by solving the resulting simultaneous algebraic equations. For each MBDoE iteration, SR proposed three candidate expressions for $r_i = f(\boldsymbol{X}, \boldsymbol{C})$ for each of the three underlying mechanisms. To simulate the PFD, r_i was substituted back into the mass and energy balance. One new experiment minimising $J(\boldsymbol{\vartheta})$, that maximised the variance in the candidate's prediction, $\widehat{\boldsymbol{\psi}}_S$, was proposed and conducted for model discrimination, expanding the dataset for the next iteration.

For each iteration of MBDoE this expanded dataset could be used to either propose a new set of equations from scratch or carry over the best-performing equations from the previous MBDoE iteration and make modifications to incorporate the new information. From a computational perspective, carrying over previously learnt knowledge is more efficient but risks carrying over biases because although SR by evolution is stochastic, expressions must pass through intermediate states to correct old misunderstandings. If the intermediate state is a poor-fitting candidate, the population of expressions may remain trapped in a suboptimal local solution. Thus, both approaches were investigated. Case Study 1A built new expressions from scratch, while Case Study 1B carried over the expressions from the previous MBDoE iteration.

Fig. 4a and 4b show the prediction MAPE for the product KPI and fitting MSE for the top three scoring expressions for each rate equation following each MBDoE iteration (I_{MBDoE}) for Case Studies 1A and 1B, respectively. For the MAPE to be small, the MSE for the selected expressions (i.e., the ones hashed) for all three rate equations had to be small.

In Case Study 1A, the MAPE decreased from 3.87% to a minimum of 0.14% by $I_{\text{MBDoE}} = 3$, while in Case Study 1B, the MAPE decreased from 3.90% to 0.30% by $I_{\text{MBDoE}} = 4$. This demonstrates that the new information from the experiments proposed by MBDoE improved the generalizability of the expressions. The proposed experiments also made sense from a model discrimination perspective. 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 promotion of better-fitting expressions.

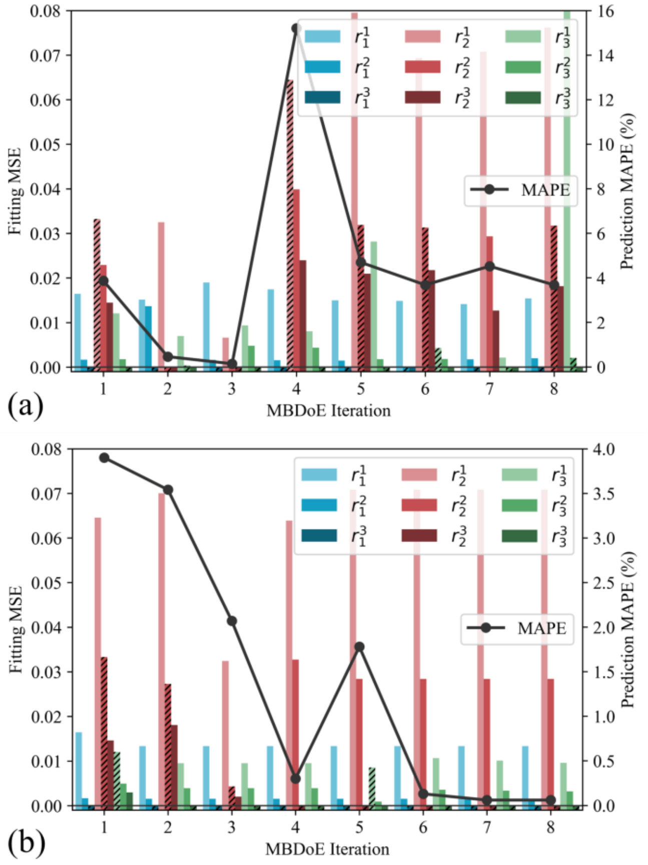


Figure 4. Mean absolute percentage error (MAPE, for prediction) and mean-square error (MSE, for fitting) for the expressions r_i^j proposed at each iteration of Case Study 1A (a) and Case Study 1B (b), where i and j denote the rate equation number and relative complexity. Bars corresponding to the top-scoring expressions used for estimating the MAPE at each iteration are hashed.

In Case Studies 1A and 1B, the MAPE eventually peaked in $I_{\text{MBDoE}} = 4$ and $I_{\text{MBDoE}} = 5$, respectively. In Case Study 1B, an approximation was proposed in $I_{\text{MBDoE}} = 5$ for r_3 that had 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 two candidates with better MSEs of 0.00001 and 0.0009, respectively. In Case Study 1B, expressions were carried over to the next MBDoE iteration and the ground truth structure for r_3 had been discovered already but had been demoted in favor of a simpler approximation. Hence, a sufficiently simple approximation can be preferred over more complex expressions, even if that more complex expression is the ground truth. This shows how the choice of metric to score and select expressions is critical. The advantage of the score definition in (1) is that expressions at 'elbow' points on the Pareto front are selected. However, other metrics will come with their own biases. Only by performing new, carefully designed experiments is it possible to reliably discriminate between different hypotheses and identify the correct model. So, when $I_{\text{MBDoE}} = 6$ added a new experiment, and the MSE of the approximation increased from 0.0086 to 0.011, the score ranking flipped, and the ground truth was re-identified.

While the MAPE recovered in Case Study 1B, it did not in Case Study 1A when SR had to propose new expressions from scratch. In Case Study 1A, the MAPE remained around 4.14% for $I_{\text{MBDoE}} \geq 5$, even as a further five unique experiments were added. This is even though SR was allowed 10 times longer to search for good expressions per MBDoE iteration in Case Study 1A compared with Case Study 1B. This result suggests that building complete expressions from scratch is more challenging. So, rather than inherited biases hindering better expression discovery, adding new information by MBDoE incrementally seems to guide SR towards the correct structure more efficiently. Therefore, hereafter, expressions were carried over from one MBDoE iteration to the next to be modified using the new information.

Thus far, in Case Studies 1A and 1B, the constraint (i.e., based on prior knowledge of state equilibration) has been active and successfully yielded expressions of the desired structure that could be easily interpreted in terms of key forward and backward driving force factors. Case Study 1C investigated the performance of the SR-MBDoE framework when lifting this constraint. Here, the MAPE decreased from 25.6% to 1.98% and plateaued for $I_{\text{MBDoE}} \geq 7$, 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.

4.2. Case Study 2: Knowledge Discovery – PFD Optimisation

Case Study 2 investigated the performance of the SR-MBDoE framework when the objective was simultaneous knowledge discovery and PFD optimisation. This was achieved by formulating $J(\boldsymbol{\vartheta})$ as a multi-objective optimisation problem with a weighting $0 \leq \alpha \leq 1$ determined automatically in Step 3 to re-balance the competition between exploration and exploitation based on the actual information and optimality gains from the previous MBDoE iteration. The aim was to hit a target final product KPI, ψ_t, within a $\kappa = \pm 3\%$ tolerance of a specific value while minimising total batch time, τ.

Table 1 shows the actual information and optimality gains, the calculated value of the weighting (i.e., α) and the total batch time for the proposed experiment for Case Study 2. Initially, MBDoE bounced between exploration (i.e., $\alpha > 0.5$) and exploitation (i.e., $\alpha < 0.5$). For as long as ΔJ_E was large, and new experiments continued to be informative, then exploration was prioritised. Once ΔJ_E became small (i.e., $\alpha = 0.2$ by $I_{\text{MBDoE}} = 5$), 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 $I_{\text{MBDoE}} = 7$ the ground truth had been discovered, so MBDoE switched to pure optimisation (i.e., $\alpha = 0.00085$). To terminate MBDoE more effectively in future, an MBDoE stopping criteria could be to iterate until $\alpha < \alpha_t$ drops below a threshold (e.g., $\alpha_t \approx 1 \times 10^{-3}$) signaling when the information gain is small enough to focus one final experiment on optimisation.

Table 1. Actual information ΔJ_E and optimality ΔJ_O gains, exploration-exploitation weighting (α) and total batch time (τ) for each MBDoE iteration (I_{MBDoE}) for Case Study 2.

I_{MBDoE}	ΔJ_E	ΔJ_O	α	τ (min)
0	-	-	0.5	100
1	0.0045	0	1	120
2	0.26	5.4	0.045	118
3	1.8	0.078	0.95	125
4	0.10	10	0.010	107
5	0.30	1.1	0.20	69
6	0.041	0.51	0.074	144
7	1.3e-06	0.0016	0.00085	83
8	8.4e-07	0.0011	0.00075	59
9	7.7e-07	0.0011	0.00066	58
10	8.5e-07	0.0012	0.00070	57

Fig. 5 shows the trajectory for the optimal PFD designed following MBDoE. The nominal total batch time was greatly reduced from $\tau = 100$ min at $I_{\text{MBDoE}} = 0$ to $\tau = 59$ min, while the final KPI satisfied the requirement, further evidencing the practical advantage and efficiency of the currently proposed SR-MBDoE digital framework. Due to the discovery of mechanistic rate expressions, it was also possible to interpret the physical trade-offs made by the optimized PFD in a way not possible for a pure machine learning approach.

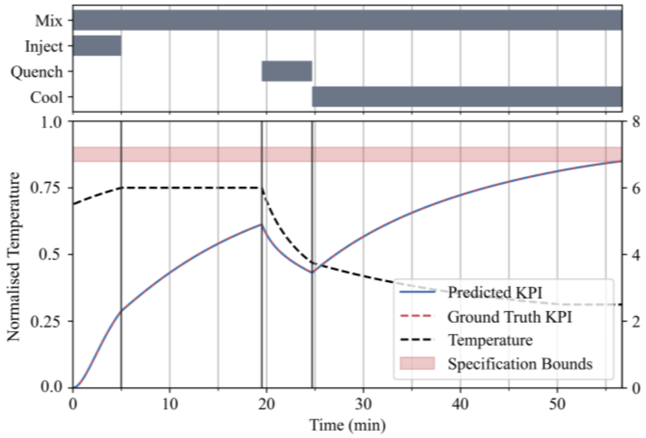


Figure 5. Predicted and actual product KPI profiles for optimized PFD parameters with model from $I_{\text{MBDoE}} = 7$ from Case Study 2.

5. CONCLUSIONS

Through two case studies, it was demonstrated that 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 by conducting carefully designed experiments was it possible to 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 synergizing 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. and Karakasidis, T.E. (2023) 'Artificial Intelligence in Physical Sciences: Symbolic Regression Trends and Perspectives', Archives of Computational Methods in Engineering, 30(6), pp. 3845–3865. Available at: https://doi.org/10.1007/s11831-023-09922-z.

Brunton, S.L., Proctor, J.L. and 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), pp. 3932–3937. Available at: https://doi.org/10.1073/pnas.1517384113.

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

Echtermeyer, A. et al. (2017) 'Self-optimisation and model-based design of experiments for developing a C–H activation flow process', Beilstein Journal of Organic Chemistry, 13(1), pp. 150–163. Available at: https://doi.org/10.3762/bjoc.13.18.

de Franca, F.O. et al. (2023) 'Interpretable Symbolic Regression for Data Science: Analysis of the 2022 Competition'. arXiv. Available at: http://arxiv.org/abs/2304.01117.

Franceschini, G. and Macchietto, S. (2008) 'Model-based design of experiments for parameter precision: State of the art', Chemical Engineering Science, 63(19), pp. 4846–4872. Available at: https://doi.org/10.1016/j.ces.2007.11.034.

Servia, M.Á. de C. et al. (2023) 'The Automated Discovery of Kinetic Rate Models -- Methodological Frameworks'. arXiv. Available at: http://arxiv.org/abs/2301.11356.