Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Hybrid Models Identification and Training through Evolutionary Algorithms

Ulderico Di Caprio[a], M. Enis Leblebici[a]*

[a] Center for Industrial Process Technology, Department of Chemical Engineering, KU Leuven, Agoralaan Building B, 3590 Diepenbeek, Belgium
* Corresponding Author: muminenis.leblebici@kuleuven.be

ABSTRACT

Hybrid modelling is widely employed in chemical engineering to generate highly accurate predictions. Such an approach merges first-principle modelling with machine learning techniques to identify and model the epistemic uncertainty from experimental data. Despite its advantages, this still requires cross-domain competencies that are difficult to find in the chemical industry and high human involvement. The possibility of automating the identification and training model would be significantly beneficial for the widespread adoption of hybrid modelling methodology within the chemical industry. This work presents a novel algorithm for the automatic identification of hybrid models (HMs) starting from the first-principle representation of the system, described by differential equation sets. The methodology formulates the problem as mixed-integer programming, identifying the equation running under uncertainty, identifying the machine learning model hyperparameters, and training the latter. The Differential Evolution algorithm drives the identification and training tasks. The methodology is validated in three cases, namely a dynamic reaction system, a dynamic bioreactor and a Lotka-Volterra oscillator deviated with polynomial or MRF equation on different levels, generating 14 validation cases. On all of them, the model correctly identifies the position of the uncertainty and the functional form to approximate it. The methodology returns automatically trained HMs with a mean absolute percentage error in the range of 10%, which is in line with the experimental error of the data. The methodology presented in this work presents a step toward the automatic generation of HMs for dynamic systems and the widespread of this technology in the chemical industry.

Keywords: hybrid modelling, automatic identification, epistemic uncertainty, differential evolution, machine learning

INTRODUCTION

Model accuracy is crucial for effective process design and optimization in the chemical industry. However, model performance often suffers from epistemic uncertainty due to unknown phenomena that current knowledge or analytical techniques cannot fully capture. This uncertainty biases predictions and compromises design quality. Hybrid modelling (HM) has emerged as a promising approach to mitigate these issues by integrating first-principle models (FPMs) with machine learning (ML) methods [1]. Here, experimental data are used to capture unmodeled phenomena in a functional form, thereby enhancing accuracy and providing insights into system behaviour.

In chemical process system engineering, FPMs are typically formulated as differential-algebraic equations that include conservation laws, thermodynamic relations, and other fundamental principles. Each of these equations may run under epistemic uncertainty, making them candidates for correction through a data-driven model (DDM). Constructing an HM requires three steps: (1) identifying the equation affected by epistemic uncertainty, (2) selecting an appropriate DDM structure to capture the unmodeled effects, and (3) training the DDM using available experimental data. The DDM must be sufficiently complex to represent the system accurately while maintaining a number of parameters that align with data availability.

Despite the benefits of HM techniques, their

widespread adoption in the chemical industry is hindered by the need for interdisciplinary expertise in chemical engineering, modelling, and ML. Other fields have benefited from automated ML frameworks that streamline model training and simplify the adoption process. An analogous automated framework for HM could facilitate its broader use, enhancing the interpretability of models and providing new insights into the system dynamics.

Currently, the identification of the epistemic uncertainty location relies heavily on domain expertise and physical insight. While some progress has been made in automating other aspects of HM, the literature lacks a comprehensive methodology for automatically locating epistemic uncertainty. Several studies have explored the selection of DDM structures and hyperparameters using approaches such as integer programming and genetic programming. For example, the ALAMO approach employs quadratic integer programming to approximate unknown physical functions by selecting suitable elementary function candidates and fitting parameters to experimental data [2]. Similarly, Narayanan et al. used genetic programming to approximate the right-hand side of differential equations, aiming to capture the entire physical relationship without additional physical constraints [3]. Furthermore, Willis and von Stosch introduced a mixed-integer linear programming framework that utilizes multivariate rational functions as approximators, validating their method on biochemical reactions and adjusting model complexity based on data availability [4].

In addition to the aforementioned techniques for generating hybrid models (HMs), recent studies have employed symbolic regression (SR) and sparse identification of dynamic systems (SINDy) for data-driven equation discovery [5, 6]. Both methodologies utilize a library of candidate functions to identify underlying equations, producing highly interpretable symbolic representations that approximate the unknown aspects of a system. However, they adopt distinct strategies tailored to different applications. Symbolic regression is typically applied to tabular data, using evolutionary algorithms to construct expression trees of elementary operations. This process identifies the optimal equation configuration that minimizes the discrepancy with experimental data. A notable advantage of SR is its ability to incorporate constraints during optimization, ensuring that the derived models are physically consistent and capable of integrating qualitative information. On the other hand, SINDy is designed for dynamic systems and focuses on uncovering the functional forms governing the temporal evolution of state variables. It employs sparse regression techniques, particularly quadratic optimization, to directly tune model parameters based on state derivative expressions without requiring the explicit solution of differential equations. While both approaches have demonstrated high performance in their respective fields, neither has been systematically evaluated for generating hybrid models in chemical processes.

To our knowledge, no framework exists that can automatically generate a hybrid model solely from an FPM representation and experimental data by simultaneously executing all three HM design steps. Addressing this gap involves solving a mixed-integer programming (MIP) problem that integrates categorical variables (e.g., the location of epistemic uncertainty), discrete variables (e.g., ML hyperparameters), and continuous variables (e.g., DDM fitting parameters). In this paper, we propose an algorithm that automatically identifies the equation affected by epistemic uncertainty, selects the most appropriate DDM hyperparameters, and trains the model using a heuristic global optimization strategy. This approach aims to enhance data efficiency while remaining compatible with the constraints of industrial data availability.

METHODOLOGY

The algorithm

The MIP problem to identify and train the HM was solved using the Differential Evolution (DE) algorithm (Scipy v.1.11.3) [7], compatible with integer variables and constraints. For the constraints, the library implements the algorithm proposed by Lampinen [8]. The MIP problem is divided into 3 subproblems, namely 1) the identification of the equation running under epistemic uncertainty, 2) the identification of the most suitable data-driven function structure to capture the said uncertainty from the training data, and 3) training the DDM identifying its continuous parameter values; these three problems are solved simultaneously by the DE. The algorithm employs polynomial equations as the data-driven function. The algorithm can activate each of the monomial terms contained in the polynomial.

The first problem is treated as a discrete optimisation problem, where the discrete variable P sets which equation in the system runs with epistemic uncertainty and should be corrected by the DDM. The discrete variable P varies between 0 and (N_s-1), where N_s is the number of equations in the system.

The second solved problem is the structure identification of the data-driven function. The algorithm implements a polynomial function as an approximator by deriving monomial terms from the states of the differential equation, constructing a K-order polynomial, and scaling the exponents of each term within the predefined range *EX_MIN* and *EX_MAX*. This ensures a balanced representation of the system dynamics, enabling accurate approximations while maintaining numerical stability and computational efficiency. This work employed K=4, EX_MIN=0.5 and EX_MAX=2. The structure of the polynomial function to train is selected, activating only certain monomial terms. In the polynomial data-driven function,

each of the monomial terms is multiplied by a continuous and a binary variable

$$M_i = \delta_i w_i \cdot f_i(x_1, x_2, \dots) \quad (1)$$

where $\delta_i \in \{0,1\}$ is the binary variable, w_i is the continuous variable, x_i is the i-th state of the differential equation system and f_i is the polynomial form related to M_i. Both δ_i and w_i of (1) are identified by the DE algorithm; the first is treated as a bounded discrete variable, while the second one is encoded as a bounded continuous variable. This work employed $w_i \in [-5,5]$. The configuration reported in (1) allows for reducing the number of parameters used in the DDM, reducing the overfitting risk in the model training phase. The number of active parameters $\delta_i w_i$ is reduced by limiting the amount of δ_i values being non-zero. Such a condition is given as a constraint to the DE

$$\sum_i \delta_i < N_{MAX}$$

where N_{MAX} is the maximum cardinality allowed by the optimisation. This work employed $N_{MAX} = 5$.

The Bayesian Information Criterion (BIC) was used as the training loss function to balance the performance of the HM with the structural complexity of the DDM and avoid overfitting the training set. The formulation by Willis and von Stosch was employed in this work [4]. The BIC was computed for each experiment, and the loss function value was calculated as their sum.

The DE algorithm was executed with 300 generations using the default settings of the SciPy implementation. Afterwards, the continuous parameters of the most performant individual were refined using the L-BFGS-B.

The computational complexity of the proposed approach is linear in the number of experiments included in the training set because the most time-consuming part is the resolution of the ODE through numerical solvers.

The test cases

Three FPMs were employed as test cases to evaluate the performance of the algorithm. These were selected to represent typical dynamic systems in the chemical engineering domain. The first test case, *the reaction system*, is a set of chemical reversible reactions described by the chemical system $A \leftrightharpoons R \leftrightharpoons S$, each running on first-order kinetic in batch conditions assumed isothermal; therefore, described through the set

$$\begin{cases} dC_A/dt = -k_{1,d} \cdot C_A + k_{1,i} \cdot C_R \\ dC_R/dt = k_{1,d} \cdot C_A - k_{1,i} \cdot C_R - k_{2,d} \cdot C_R + k_{2,i} \cdot C_S \\ dC_S/dt = k_{2,d} \cdot C_R - k_{2,i} \cdot C_S \end{cases}$$

with $k_{1,d} = 0.3\ min^{-1}$, $k_{1,i} = 0.1\ min^{-1}$, $k_{2,d} = 0.02\ min^{-1}$, $k_{2,i} = 0.01\ min^{-1}$. The equations were solved with initial conditions in the range [0, 10], with 10 points obtained using the Latin hypercube sampling (LHS).

The second test case is *the Lotka-Volterra oscillator*, described by the system

$$\begin{cases} dx/dt = (1-y) \cdot x \\ dy/dt = (x-1) \cdot y \end{cases}$$

The equations were solved with initial conditions in the range [0, 1], with 10 points obtained using the LHS.

The third evaluation case is *the bioreactor system*, describing the growth of biomass, X, in a continuous bioreactor fed with the nutrient S, influenced by Monod kinetics. The system is described by the equation set

$$\begin{cases} \dfrac{dX}{dt} = \dfrac{\mu_{max} \cdot S}{K_S + S} \cdot X - D \cdot X \\ \dfrac{dS}{dt} = D \cdot (S_{in} - S) - \dfrac{\mu_{max} \cdot S}{K_S + S} \cdot \dfrac{X}{Y_{xs}} \end{cases}$$

with $\mu_{max} = 0.4\ min^{-1}$, $K_S = 0.5\ mol/L$, $S_{in} = 1\ mol/L$, $Y_{xs} = 0.6\ g/mol$ and $D = 0.1\ min^{-1}$. The equations were solved with initial conditions in the range [0.2, 1], with 10 points obtained using the LHS.

Each of the equations in the FPMs was deviated using 2nd-order polynomial or 1st-order MRF functions to mimic the epistemic uncertainty source. The employed parameters were sampled in the range [-1,1]; following that, the output was multiplied by a constant value to enhance the effect of the deviation.

For each validation case, one of the abovementioned deviations was applied to one of the set equations individually. Therefore, the algorithm was validated using 14 cases combining the FPMs and the deviation position. These served as experimental data for the algorithm to identify the position of the epistemic uncertainty and the most suitable data-driven function structure for the HM generation.

Each of the deviated equations was solved using LSODA solver (SciPy v1.11.3). Each of the experiments contained 14 points uniformly sampled in the time range [0,20] min; noise of ±10% was added to the output, mimicking the measurement noise. The test set for the reaction system comprises 5 train points and 5 test points randomly selected. The test set for the Lotka-Volterra oscillator and the bioreactor system comprises 4 points selected to be within the initial conditions convex-hull.

The trained models were evaluated on a test set. The quantitative comparison of the trained HM performance was executed using 1) the coefficient of determination (R^2), to evaluate the explained variance of the model, 2) the mean absolute percentage error (MAPE) to have a quantitative deviation of the model from the point weighted on their experimental value, and 3) the mean absolute percentage error (MAE) to have a quantitative deviation of the model prediction from the experiments.

RESULTS

The results of the identification algorithm are reported differentiated by the test case on which they have

been validated. Tables 1, 2 and 3 report the accuracy of the identified HM on the overall test set, the actual level of the epistemic uncertainty within the equations and the one identified by the algorithm for the reaction system, the Lotka-Volterra oscillator and the bioreactor system, respectively. Figures 1, 2, and 3 report the graphical representation of the model performance on selected experiments contained in the test set, respectively, for the reaction system, the Lotka-Volterra oscillator, and the bioreactor system.

The reaction system

In the reaction system, the developed algorithm correctly identifies the epistemic uncertainty location within the equations set for all the proposed test conditions (Table 1). Moreover, HMs identified from the algorithm return low prediction errors on the test set.

Table 1. Quantitative performance of the identified HM from the algorithm on the test set. The table reports the equation on which the deviation was applied, the one identified by the algorithm, and the obtained metrics.

Actual		*Identified*			
Level	Deviation	Level	MAPE	MAE	R^2
1	Poly	1	10.8%	0.48	0.11
	MRF	1	18.4%	10.53	0.59
2	Poly	2	15.7%	7.32	0.82
	MRF	2	7.55%	0.43	0.24
3	Poly	3	6.07%	0.27	0.29
	MRF	3	6.3%	0.56	0.77

Excluding the system deviated on the 1st equation with MRF and the system deviated on the 2nd equation with polynomial, the trained HM predicted the system behaviour with a MAPE error in line with the given experimental error (i.e., ± 10%). The quality of the model is also confirmed by the MAE value, returning low metrics; this highlights how the error is evenly distributed within the entire range without any significant bias toward any area of the output range. On the other hand, in the two cases with high MAPE and MAE, the model cannot accurately predict the output in some experiments contained in the training set. The low amount of data mainly drives this behaviour for a system having 3 states. Performing the convex-hull selection also in this case, the train set comprises 8 points and improved performances; however, this was not included because of poor reliability on the test set.

Despite the low MAPE and MAE values, the R^2 obtained on the test set is quite low for some of the cases (e.g., the one with polynomial deviation on the 1st equation or the MRF deviation on the 2nd equation), showing that the trained model has limited capabilities in explaining the variance of the experimental set. We hypothesise that this is driven by the poor ability to identify continuous parameters. The algorithm uses a heuristic approach to identify all the system parameters, but it is quite inefficient for continuous parameters since heuristics do not guarantee the minimisation of the loss function for a given configuration. On the other hand, the continuous parameters can be identified by gradient-based optimisation approaches, improving the quality of the training.

From Figure 1, it is possible to observe how the employed FPM has significant deviations compared to the experimental data, especially for the profile of the reactant S, where it predicts a completely different behaviour. On the other hand, this effect is mitigated by including DDM, which corrects the model prediction and correctly identifies the deviation source. This results in improved model metrics, such as the MAPE decrease of 5 folds. The case reported in Figure 1 runs with epistemic uncertainty on S concentration. S is being generated from something else that is not included in the mathematical system. The mechanistic model of the system closes the mass balance, but in reality, the system description violates this hypothesis. To ensure mass balance, a further investigation of what is causing the S growth is required.

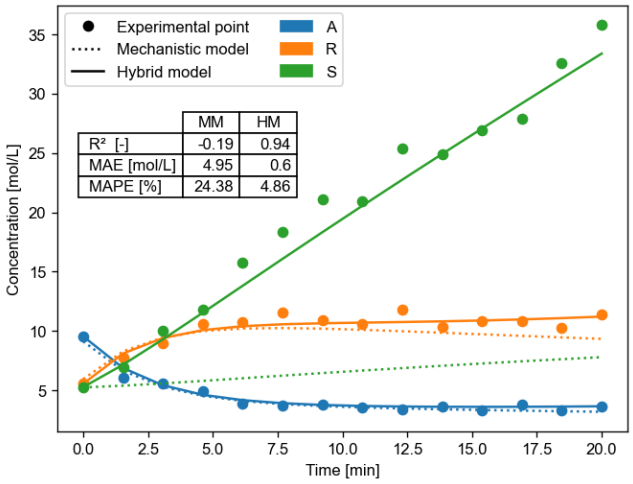

Figure 1. First-principles model and hybrid model performance on one of the test examples on the reaction system. The reported experiment is obtained with MRF deviation on the 3rd equation.

The Lotka-Volterra oscillator

Like the reaction system, also for the Lotka-Volterra case, the developed algorithm performs correct identification of the epistemic uncertainty location both for polynomial and MRF deviation (Table 2). The metrics reported in Table 2 show the enhanced prediction accuracy of the identified HM. For all the validation cases, the HM has low MAPE and MAE with R^2 very close to 1, showing how the HM was able to explain the variance of the experimental system completely.

Figure 2 depicts the performance of the mechanistic and HM on one of the test set experiments, having deviation on the 2nd equation with a polynomial function. Here, it is possible to observe how the FPM shows high

deviation on both the variables with error accumulating with time, resulting in a significant shift of the actual and modelled peaks for $t > 7.5\ min$. Additionally, the FPM returns a wrong prediction of peaks amplitude, with a systematic overestimation for both the system states. This behaviour is not observed with the identified HM, returning accurate predictions within the entire simulation time. The HM has effects not only on the peak position but also on their magnitude, which are now correctly predicted. In summary, the HM correctly identifies the behaviour of the test system returning MAPE around 5% and an R^2 = 0.98, where the FPM has worse prediction accuracy with an R^2=-0.05, indicating that the model returning the average of the points return better prediction than the FPM, and a MAPE around 38%, with a value 7 times higher than the value obtained with the HM predictions.

Table 2. Quantitative performance of the identified HM from the algorithm on the test set. The table reports the equation on which the deviation was applied, the one identified by the algorithm, and the obtained metrics.

Actual		*Identified*			
Level	Deviation	Level	MAPE	MAE	R^2
1	Poly	1	5.29%	0.066	0.96
	MRF	1	5.51%	0.068	0.95
2	Poly	2	5.68%	0.051	0.96
	MRF	2	8.95%	0.084	0.92

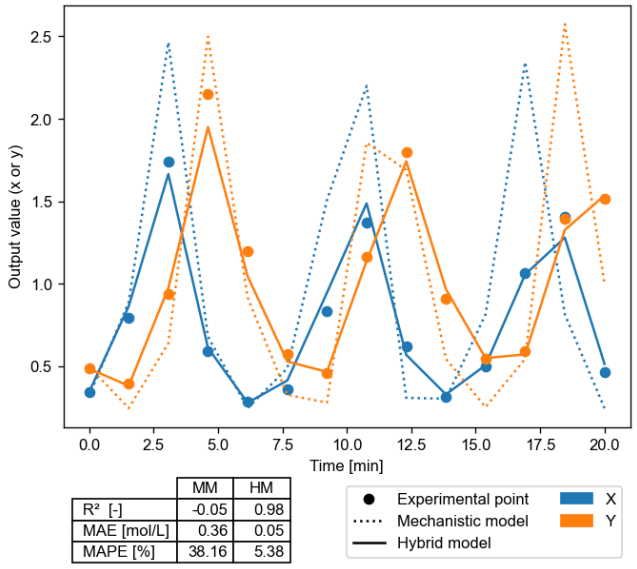

Figure 2. First-principles model and hybrid model performance on one of the test examples on the Lotka-Volterra oscillator. The reported experiment is obtained with polynomial deviation on the 2nd equation.

The bioreactor system

When describing the bioreactor system, the model correctly identifies all the validation cases, with deviations both on biomass growth and substrate profile (Table 3). The source of the epistemic uncertainty has been correctly identified by the algorithm, as well as the approximation for the deviation function. The algorithm did not show any identification bias (i.e., model accuracy or level identification) among the various cases, returning for each of the test case metrics with the same order of magnitude.

Table 3. Quantitative performance of the identified HM from the algorithm on the test set. The table reports the equation on which the deviation was applied, the one identified by the algorithm, and the obtained metrics.

Actual		*Identified*			
Level	Deviation	Level	MAPE	MAE	R^2
1	Poly	1	6.02%	0.027	0.77
	MRF	1	5.4%	0.033	0.88
2	Poly	2	6.87%	0.042	0.8
	MRF	2	6.00%	0.035	0.83

Figure 3 depicts one of the test experiments for the case having deviation on the 1st equation described by a polynomial function. Here, the FPM shows significant deviation that is cumulating over time; such behaviour is related to the fact that the numerical resolution of differential equations relies on predicted system state values to run the prediction at the current state, letting the error accumulate over time.

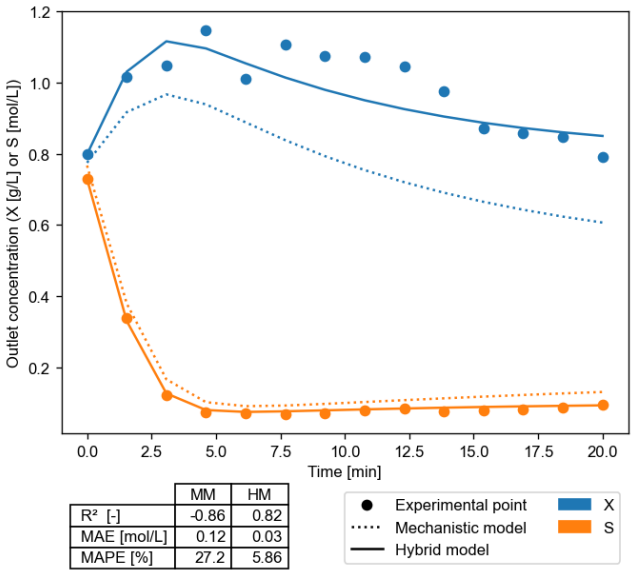

Figure 3. First-principles model and hybrid model performance on one of the test examples on the biochemical system. The reported experiment is obtained with polynomial deviation on the 1st equation.

On the other hand, the HM prediction correctly overlaps the experimental points within the entire time range without any significant deviation happening over time, demonstrating the robustness and reliability of the HM obtained with the proposed methodology. It is important to note that both the FPM and the HM identify the same

overall trend, but the HM better identify the change in system behaviour (e.g., the location of the maxima for the X profile), the magnitude of the first derivate of the profile and the plateau location for both the curves. Such an improvement in behaviour prediction resulted in better metrics for the HM over the FPM; the FPM returned a MAPE=27.2% while the R^2=-0.86, quantitively proving the poor prediction capabilities of the FPM. On the other hand, the HM results in a MAPE=5.86%, a value around 4.6 times lower than the one obtained for the FPM, and an R^2=0.82.

ALTERNATIVE APPROACHES

A primary limitation of the proposed approach is the need for iterative ODE resolution during parameter identification. This process significantly prolongs identification time and is susceptible to mathematical instabilities, thereby reducing compatibility with derivative-based optimizers. A potential solution to mitigate these issues is the application of SINDy-like techniques [6]. This alternative framework would allow the reformulation of the identification problem using quadratic integer programming to locate epistemic uncertainty while also accommodating multiple model deviations.

CONCLUSIONS

This work presents an automatic algorithm that identifies and trains hybrid models for dynamic chemical processes. Starting from a first-principle model under epistemic uncertainty, the method integrates missing information from experimental data into a data-driven component. It locates the epistemic uncertainty within the system equations and selects the most suitable data-driven model (DDM) for its mitigation. A constrained mixed-integer Differential Evolution algorithm is used to determine the uncertainty's position along with the hyperparameters and parameters of the DDM.

To work with limited data and minimal experimentation, the method employs a sum of monomial terms as the data-driven function and uses the Bayesian information criterion as the loss function, thereby reducing model complexity and overfitting risk. The approach was validated on a reactive chemical system, a bioreactor with biomass growth, and the Lotka-Volterra oscillator, accurately identifying uncertainty and achieving prediction errors (MAPE) below 10% on unseen cases.

This methodology marks a significant step toward automating hybrid model creation in the chemical industry. Future improvements include integrating deterministic algorithms for continuous parameter identification, testing additional chemical cases, and extending uncertainty identification to model parameters beyond differential equations.

ACKNOWLEDGEMENTS

The authors acknowledge funding from the KU Leuven project "HyPro - Automatic hybrid digital twins for process modelling" (C3/23/007).

REFERENCES

1. Schweidtmann AM, Zhang D, Von Stosch M. A review and perspective on hybrid modeling methodologies. *Digit. Chem. Eng.* 10:100136 (2023). doi.org/10.1016/j.dche.2023.100136
2. Wilson ZT, Sahinidis NV. The ALAMO approach to machine learning. *Comput. Chem. Eng.* 106:785–95 (2017).doi.org/10.1016/j.compchemeng.2017.02.010
3. Narayanan H, Bournazou MNC, Gosálbez GG, Butté A. Functional-Hybrid modeling through automated adaptive symbolic regression for interpretable mathematical expressions. *Chem Eng J.* 430:133032 (2021). doi.org/10.1016/j.cej.2021.133032
4. Willis MJ, Von Stosch M. Simultaneous parameter identification and discrimination of the nonparametric structure of hybrid semi-parametric models. *Comput. Chem. Eng.* 104:366–76 (2017). doi.org/10.1016/j.compchemeng.2017.05.005
5. Angelis D., Sofos F., Karakasidis T.E. Artificial Intelligence in Physical Sciences: Symbolic Regression Trends and Perspectives. Arch Computat Methods Eng 30, 3845–3865 (2023). doi.org/10.1007/s11831-023-09922-z
6. Brunton SL., Proctor JL., Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci. 113:3932-3937 (2016). doi.org/10.1073/pnas.1517384113
7. Storn R, Price K. Differential Evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* 11(4):341–59 (1997). doi.org/10.1023/a:1008202821328
8. Lampinen J. A constraint handling approach for the differential evolution algorithm. Vol. 2, *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*. 2003. doi.org/10.1109/cec.2002.1004459