Scientific Reports

https://doi.org/10.1038/s41598-026-38090-w

Article in Press

Interpretable machine learning for optimized dimethyl ether production from bio-methanol

Received: 25 July 2025

Accepted: 28 January 2026

Published online: 19 February 2026

Cite this article as: Mokari M., Rahmani M. & Atashrouz S. Interpretable machine learning for optimized dimethyl ether production from bio-methanol. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-38090-w

Mohsen Mokari, Mohammad Rahmani & Saeid Atashrouz

Interpretable Machine Learning for Optimized Dimethyl Ether Production from Bio-Methanol

Mohsen Mokari, and Mohammad Rahmani*, Saeid Atashrouz

Chemical Engineering Department, Amirkabir University of Technology, Tehran, Iran

CleanTech Research Laboratory, Amirkabir University of Technology, Tehran, Iran

* **Corresponding author:** Email: m.rahmani@aut.ac.ir, Tel: (+98)-21-64543198, Fax: (+98)-21-66405847

Abstract

Hybrid models often referred to as gray-box models offer a promising approach by combining the flexibility of data-driven techniques with the accuracy and physical interpretability of first-principles models. This study evaluates a range of mathematical modeling techniques in the context of chemical reaction engineering, with a focus on the production of dimethyl ether (DME) from bio-methanol in a fixed-bed reactor. A comprehensive case study was conducted, beginning with the development of a first-principles model to solve a system of governing equations and generate 7,000 synthetic data points with added noise. Three black-box machine learning algorithms K-Nearest Neighbors (KNN), Gradient Boosting Regressor (GBR), and eXtreme Gradient Boosting (XGB) were employed for predictive modeling. In parallel, hybrid modeling approaches were developed to estimate reaction rates and correct reactor outputs. Model performance was assessed using metrics such as Mean Squared Error (MSE) and the coefficient of determination (R^2), using key variables including inlet molar flow rate, initial temperature, pressure, and the outlet concentrations of methanol, dimethyl ether, and water, as well as overall conversion. Results indicated that the data-driven models performed exceptionally well, with hybrid models offering

comparable accuracy while maintaining interpretability. Finally, process optimization was performed using the eXtreme Gradient Boosting model integrated with a Differential Evolution algorithm. The optimized operational conditions achieved a high dimethyl ether conversion rate of 84.3%, with a minimal temperature rise of 84.9 K.

Keywords: Hybrid Modeling, Machine Learning, Long Short-Term Memory, Differential Evolution

1.Introduction

Since the Third Industrial Revolution, modeling has played a central role in process monitoring, control, optimization, and design within the chemical industry [1]. With the onset of the Fourth Industrial Revolution and the rise of large-scale industrial data, data-driven modeling frameworks have gained prominence [2]. Nevertheless, mechanistic models remain indispensable due to their interpretability and strict adherence to physical laws [3]. Recent research emphasizes the integration of domain knowledge with machine learning to develop reliable and explainable AI-based models [4]. Selecting an appropriate hybrid modeling structure depends on system characteristics and the reliability of available information [5,6]. Hybrid approaches typically begin with assessing the mechanistic model; parallel configurations are used when significant model–data discrepancies exist [7,8], while serial configurations are preferred when the first-principles model is sufficiently accurate [9]. In serial architectures, the mechanistic component enforces physical constraints, while the data-driven part captures complex or poorly understood behaviors such as reaction kinetics [1–3,5–9].

The inherent complexity of chemical systems and their interdependent physical phenomena have historically made first-principles modeling challenging [10,11]. While machine learning provides powerful tools for handling highly nonlinear systems [12], mechanistic models grounded in conservation laws, thermodynamics, and kinetics remain valuable for their physical consistency [13,14]. These models are implemented through equation-oriented strategies [15,17] or physics-based formulations that automate model generation based on system geometry and assumptions [18]. Comprehensive mechanistic models often require auxiliary correlations, such as vapor–liquid equilibrium, kinetic models, transport properties, and packing behavior [19,20]. In contrast, black-box models rely solely on data and thus require large, high-quality datasets to achieve acceptable performance [21,22].

Hybrid (gray-box) models have emerged as a promising solution, combining the interpretability of mechanistic models with the flexibility of data-driven techniques [23,24]. These models improve predictive accuracy, reduce data requirements, and maintain physical consistency [25]. Common hybrid structures include serial configurations where mechanistic and data-driven components are linked sequentially and parallel architectures, where the data-driven model compensates for mechanistic discrepancies [26,27]. Serial models are particularly effective when the mechanistic model captures major dynamics but requires support in modeling complex sub-phenomena [28], whereas parallel models are preferred when the mechanistic model exhibits substantial deviations from empirical data [29–32]. Parallel frameworks are particularly useful for systems dominated by nonlinearities or unmodeled dynamics [33,34], although their performance is limited to the range of the training data [35]. Serial structures, by contrast, can leverage simulated datasets generated by first-principles models when experimental data are scarce [36,37].

Beyond serial and parallel architectures, physics-informed models introduce physical constraints directly into the learning process through the cost-function gradients, as demonstrated in physics-informed neural networks (PINNs) [38,39]. The development of hybrid models typically involves defining modeling objectives, establishing conservation laws, identifying uncertain components, selecting the appropriate machine-learning strategy, and training the model on available data [40].

Hybrid models offer advantages such as improved extrapolation, reduced dimensionality, and the ability to incorporate new data efficiently [16]. They are often easier to construct than detailed mechanistic models while retaining adaptability and interpretability [41]. However, their performance is ultimately constrained by inaccuracies or simplifications in the underlying mechanistic structure [42–44]. Thus, expert knowledge remains essential for

selecting an appropriate hybrid architecture and effectively integrating mechanistic and data-driven components [1–3,5–9].

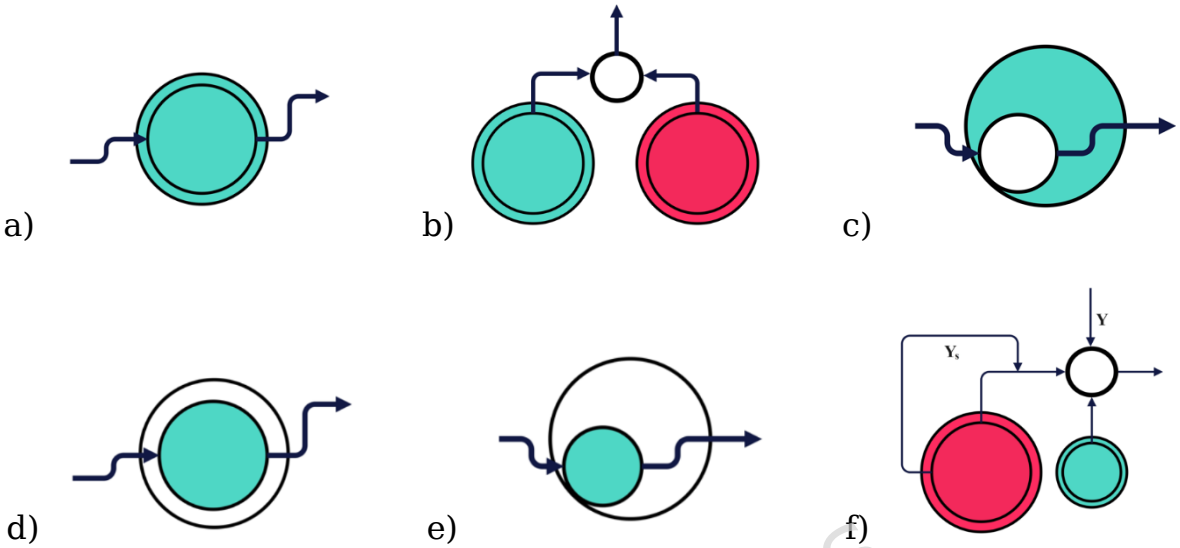


Fig.1. Schematic of various hybrid modeling methods: Surrogate (a), Correction (b), Estimation (c), Physics-Informed (d), Structural (e), and Calibration (f) [26]

First-principles modeling relies on solving complex algebraic or partial differential equations, often requiring analytical or numerical techniques [45]. Although data-driven models offer flexibility, they typically demand large datasets and exhibit limited extrapolation capability compared with mechanistic approaches [46]. Hybrid or gray-box models overcome these limitations by integrating white-box and black-box components to exploit their complementary strengths [47], using either sequential or parallel structures [48]. Several studies have demonstrated the effectiveness of hybrid approaches in catalytic reactor modeling and process prediction. Lou et al. [42] developed a hybrid model for ethylene oxidation over a silver catalyst by combining first-principles descriptions of a fixed-bed reactor with support vector regression to predict catalyst deactivation using industrial operational data. Similarly, another work [43] proposed a hybrid empirical framework for monitoring and forecasting catalyst lifetime in industrial reactors by integrating an adiabatic first-principles model with a partial least squares model. Additional research has focused on hybrid residual modeling,

where subspace identification techniques were used to capture discrepancies between plant data and mechanistic predictions, improving accuracy in continuous stirred tank reactors [51]. Hybrid models coupling first-principles frameworks with artificial neural networks have also shown significant potential in describing catalyst deactivation in fixed-bed units [54], advancing CO_2 conversion prediction [56], and improving kinetic modeling accuracy and robustness for dimethyl ether synthesis [57].

Building upon these developments [42–57], the present work introduces a unified and interpretable hybrid modeling architecture that integrates mechanistic simulation, data-driven kinetic estimation via ensemble learning, sequential extrapolation using Long Short-Term Memory (LSTM) networks, and optimization through Differential Evolution. Whereas prior studies typically coupled first-principles models with individual machine-learning elements to target specific tasks such as catalyst deactivation [42–44], temperature-field prediction [45,46], or emission estimation [47,48], the proposed framework establishes a generalizable hybrid structure that adaptively corrects mechanistic predictions and infers unknown kinetic relationships without predefined rate laws. This dynamic correction–estimation strategy preserves physical interpretability by embedding first-principles constraints while ensuring transparent model adjustments, thereby improving reliability under non-ideal operating conditions. Accordingly, the primary objective of this study is to develop an interpretable, optimization-integrated hybrid model for DME synthesis in a fixed-bed reactor that combines mechanistic simulation, data-driven kinetic inference, LSTM-based extrapolation, and Differential Evolution–based optimization.

Furthermore, by coupling the developed hybrid models with a differential evolution–based optimization module, the framework bridges hybrid process modeling with optimization-driven design, enabling direct evaluation of reactor performance while maintaining computational efficiency, consistency, and high interpretability. Accordingly, the proposed work

advances beyond existing hybrid frameworks by offering a comprehensive, physically interpretable, and optimization-integrated modeling structure for dimethyl ether production from bio-methanol. This study aims to develop an interpretable hybrid modeling framework integrating first-principles and data-driven methods for optimized DME synthesis. Unlike prior surrogate-based works, this framework explicitly combines mechanistic constraints with data-driven correction and optimization modules. All implemented models in this study were executed using the Python programming language in the Google Colab environment, with libraries including PyTorch [17], TensorFlow [18], Pandas [19], NumPy [20], Scikit-learn [21], and SciPy [22]. (The overall Schematic of the overall modeling and optimization workflow is illustrated in Fig.2.)

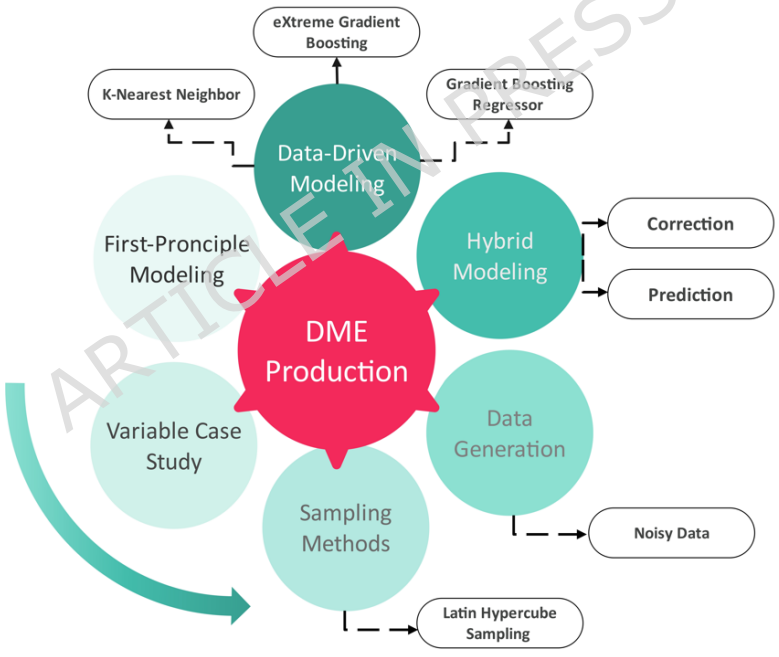


Fig.2. Schematic of the overall modeling and optimization workflow, illustrating first-principles data generation, and hybrid/data-driven model development (KNN, GBR, XGB, LSTM) for DME reactor performance.

2. Materials and Methods

2.1. Artificial Neural Networks

Artificial neural networks (ANNs) were implemented to capture complex nonlinear relationships between reactor inputs and outputs. In this study, both feedforward and recurrent architectures were examined, with model structures and hyperparameters optimized based on prior benchmark studies Model training employed the mean squared error as the loss function and the Adam optimizer, which consistently yielded the best convergence performance in previous studies [10,60].

$$w_{new} = w_{old} - \alpha \frac{\partial(\text{loss function})}{\partial w} \tag{8}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{9}$$

2.2. Machine Learning Algorithms

2.2.1 Ensemble and Boosting Methods

Ensemble learning combines multiple base estimators to improve robustness and predictive accuracy in machine learning models [61]. Among ensemble strategies, boosting methods notably Gradient Boosting Regressor (GBR) and eXtreme Gradient Boosting (XGB) have shown superior performance in nonlinear regression problems, particularly in chemical process modeling.

The Gradient Boosting Regressor iteratively builds an additive model by sequentially training weak learners to minimize residual errors from previous iterations, resulting in progressively refined predictions [61]. XGBoost, introduced by Chen and Guestrin [64], extends this framework by incorporating L_1 and L_2 regularization, second-order gradient information, and efficient handling of missing data, thereby enhancing generalization, computational efficiency, and resistance to overfitting. In this study, GBR and XGB were integrated into the hybrid modeling framework to capture complex nonlinear relationships between input variables and reactor outputs. Hyperparameters such as learning rate, tree depth, and number of estimators

were optimized via cross-validation to minimize prediction error and improve model generalization [61,64].

$$f_n(x_i) = f_0 + \alpha \,.\, T_1(x_i) + \alpha \,.\, T_2(x_i) + \ldots + \alpha \,.\, T_n(x_i) \tag{10}$$

For XGBoost, node splitting decisions in each tree are guided by the Gain parameter, which quantifies the improvement in the objective function. A negative gain indicates that splitting the node is not beneficial, and thus the node should be pruned. The Gain is computed as follows:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma, \quad I_L \cup I_R = I \tag{11}$$

2.2.2 K-Nearest Neighbors Algorithm

The K-Nearest Neighbors (KNN) algorithm is a non-parametric method that does not construct an explicit model; instead, it relies on storing training examples to make predictions. In regression tasks, KNN identifies the K nearest neighbors of a query point and estimates the target value as the average of these neighbors. The procedure involves selecting the parameter K and computing distances such as Euclidean, Manhattan, or Minkowski between the query and training samples [61–63, 65].

$$\text{Euclidean Distance } d(x,y) = \sqrt{\sum_{1}^{n}(x_i - y_i)^2} \tag{12}$$

$$\text{Manhattan Distance } d(x,y) = \sum_{1}^{n}|x_i - y_i| \tag{13}$$

$$\text{Minkowski Distance } d(x,y) = \sum_{1}^{n}(|x_i - y_i|^p)^{1/p} \tag{14}$$

The selected surrogate models (KNN, GBR, and XGB) were chosen because the dataset represents steady-state reactor samples rather than temporal sequences. Each data point corresponds to a specific reactor position and set of operating conditions; therefore, the problem is non-temporal in nature. Accordingly, tree-based ensemble methods (GBR and XGB) and instance-based learning (KNN) were selected for their strong ability to capture nonlinear steady-state relationships between the input and output variables without relying on sequential dependencies. These algorithms also offer computational efficiency, robustness to overfitting, and ease of interpretation compared to deep sequence-aware architectures such as LSTMs or Temporal Convolutional Networks (TCNs), which are more suitable for dynamic or time-series data. Furthermore, to complement these static surrogates, an LSTM-based hybrid correction framework (Section 3.2.1) was employed specifically for spatial extrapolation along the reactor length, where sequential dependencies become relevant. This hybrid design ensures that each modeling approach is utilized within its most appropriate context.

2.3. Recurrent Neural Networks

Recurrent neural networks (RNNs) are designed to process sequential data by incorporating feedback connections that allow information from previous steps to influence current predictions [10,66]. However, conventional RNNs often struggle to capture long-term dependencies due to gradient vanishing or exploding issues. To overcome this limitation, the Long Short-Term Memory (LSTM) architecture was employed. LSTMs introduce gated memory units that selectively retain or discard information through forget and update mechanisms, enabling the network to learn long-range dependencies effectively [66,67].

In this paper, the LSTM model was applied to represent sequential spatial correlations along the reactor length, allowing accurate extrapolation of

process variables where conventional feedforward networks fail to capture cumulative dependencies.

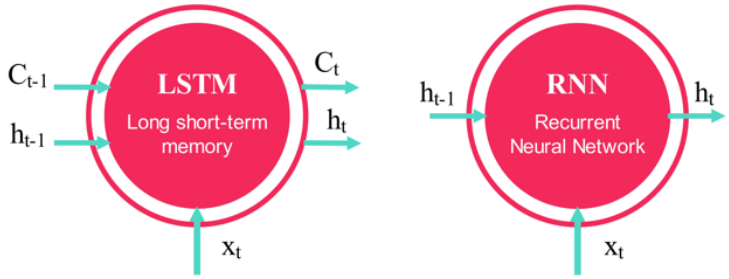


Fig.3. Recurrent vs Long Short-Term Memory neural networks

2.4. Sampling Methods

In this work, we introduce the concept of intelligent sampling [68], which entails the systematic design or refinement of algorithms tailored to specific sampling objectives. Through both qualitative and quantitative evaluations, it becomes evident that several relatively simple algorithms can be adapted effectively for tasks such as surrogate modeling, hyperparameter optimization, and data analysis. These modified algorithms often outperform more complex alternatives currently in practice, achieving higher efficiency in terms of both time and computational resources.

Based on a comprehensive review, this study employs the Latin Hypercube Sampling (LHS) method, a widely adopted approach for generating initial sample points in surrogate and data-driven modeling. The core principle of LHS is to divide each input dimension d into N intervals and then select a coordinate randomly or at the interval center along each dimension. This procedure yields N sample points in d-dimensional space. LHS is computationally efficient and capable of generating any number of samples while maintaining favorable visual properties, as the samples in each dimension uniformly span the corresponding range. Nevertheless, it should

be noted that LHS does not inherently guarantee uniform coverage of the entire design space [68,70,71].

In machine learning–based modeling, data acquisition and utilization are critical, and the choice of sampling strategy directly affects model performance. Sampling allows for the selection of representative data points from a dataset with unknown parameters, offering the advantages of reduced cost and expedited collection. Effective sampling requires careful consideration of objectives, variation ranges, and the required number of samples. In this study, four sampling strategies are investigated: random sampling, LHS [68], Halton, and Sobol [70]. In all three random sampling approaches, each element of the dataset has an equal probability of selection. The Halton and Sobol methods, in contrast, are quasi-random sampling techniques designed to generate low-discrepancy sequences, enhancing coverage efficiency in high-dimensional spaces [71].

Table.1. Comparison of different sampling methods in terms of various adaptabilities [68]

	Sampling Methods					
Adaptability to	Uniform	Grid	LHS[1]	Piosson Disk	GreedyFP	BC[2]
Arbitrary number of samples	+	×	+	×	+	+
Randomness	+	+	+	+	+	+
Probability density function	+	×	+	×	+	+

[1] Latin Hypercube Sampling
[2] Best Candidate

Modified domain extents	+	×	×	+	+	+
High dimensional sampling	+	+	+	+	+	+
Large number of samples	+	+	+	+	+	+

2.5. Data Preparation

The databank of 7000 runs was generated by the authors using the validated first-principles model of the fixed-bed DME reactor. Input variables were sampled using the Latin Hypercube Sampling (LHS) method to cover the full range of operating conditions, and the corresponding reactor outputs were obtained by solving the governing conservation equations. The resulting dataset served as the basis for training the data-driven and hybrid modeling frameworks. To introduce realistic variability, the input variables were perturbed using a pseudo-random binary sequence (PRBS), and Gaussian white noise with an amplitude between 5% and 10% of the corresponding signal was added to the outputs to simulate experimental uncertainty. We normalize the data using various statistical functions to standardize the range of data variations, making the model training process easier and faster. Here, the data is normalized using the Min-Max Scaler method, which scales the data to a range between 0 and 1 as shown in e.q.15 .Additionally, two other common methods for normalizing data are the L_1 and L_2 norm [62].

After the data normalization stage, we separate the features that serve as the model's input and output, and then divide the data into three categories: training data (65%), validation data (10%), and test data (25%).

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}} \tag{15}$$

$$L_1 \text{ norm (Manhattan Distance)} = \frac{x_{ij}}{|x_1| + |x_2| + |x_3| + \ldots + |x_n|} \tag{16}$$

$$L_2 \text{ norm (Euclidean Distance)} = \frac{x_{ij}}{\sqrt{x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2}} \quad (17)$$

2.6. Modeling the Dimethyl Ether Reactor

The core of a large-scale dimethyl ether plant is a chemical reactor filled with acidic catalysts like gamma-alumina (γ-Al_2O_3), which acts as a conventional methanol-to-dimethyl ether converter through the highly exothermic dehydration reaction shown e.q.15 [59].

$$2CH_3OH \overset{K_{eq}}{\leftrightarrow} CH_3OCH_3 + H_2O \qquad \Delta H_0 = -23.4 \frac{kJ}{mol} \quad (18)$$

γ-Al_2O_3 at the heart of a large-scale dimethyl ether (DME) production facility lies a catalytic reactor packed with acidic materials such as gamma-alumina (γ-Al_2O_3), which functions as a conventional catalyst for the conversion of methanol to DME through a highly exothermic dehydration process, as represented in Equation 15. Gamma-alumina is widely employed in this application due to its favorable properties, including high surface acidity, stability, cost-effectiveness, ease of fabrication, and widespread availability. This catalyst demonstrates satisfactory performance under a broad range of operating conditions, typically between 523 and 673 K in temperature and 10 to 70 bar in pressure. Interestingly, successful operation at relatively lower temperatures has also been demonstrated in similar prior research efforts.

Two principal reactor configurations are considered for DME synthesis: an adiabatic reactor and a water-cooled reactor. The adiabatic design retains the heat generated during the reaction, leading to a rise in temperature that accelerates the reaction rate. Conversely, in the water-cooled design, heat is extracted via a surrounding water jacket containing saturated water, which helps shift the chemical equilibrium toward increased DME production. Prior to implementing machine learning techniques or hybrid modeling

approaches, it is essential to construct a reliable mathematical model that reflects the actual system behavior and serves as a foundation for generating the required dataset. This model incorporates mass, energy, and momentum conservation principles, initial and boundary conditions, reaction rate laws, kinetic expressions, physical property correlations, and an appropriate equation of state specifically, the Peng–Robinson model along with several critical modeling assumptions, as outlined below:

- The reaction mixture is homogeneous and takes place entirely in the gas phase.
- The system is considered to operate at steady state.
- Radial gradients are ignored, and the reactor is modeled in one dimension along its length.
- Due to high gas velocities, axial dispersion of mass and energy is negligible.
- A plug flow regime is assumed throughout the reactor.
- Pressure drop is estimated using the Ergun equation.
- The reactor employs a fixed-bed configuration with defined porosity.
- Non-ideal gas behavior is accounted for using the Peng–Robinson equation of state to compute fugacities and density.
- Lateral heat losses through the reactor wall are considered negligible due to insulation.

Because of the high gas velocity and small catalyst particle size, convective heat and mass transport dominate, justifying the plug flow assumption. Furthermore, intraparticle resistance is considered negligible, allowing for the assumption of uniform concentration and temperature within catalyst pellets. The reactor's high length-to-diameter ratio (e.g., 2.02 and 80.8) further supports the neglect of radial variations, reinforcing the use of a one-dimensional axial model. These assumptions are commonly adopted in previous research and are considered well-established.

The mathematical model governing the process consists of a system of first-order ordinary differential equations with inlet boundary conditions. These primary equations are supplemented with auxiliary expressions for physical properties, reaction kinetics, and thermodynamic behavior. The reaction kinetics play a pivotal role in determining model accuracy, particularly the reaction rate expression (R_i). In this study, among several kinetic models examined, the second kinetic expression (referred to as "r") was chosen for further analysis due to its superior alignment with experimental observations. This model is based on equilibrium constant formulations and offers higher predictive accuracy. Accurate determination of reaction kinetics is crucial, especially when modeling the methanol dehydration reaction catalyzed by γ-Al_2O_3, as it directly impacts the reliability of the generated dataset used for machine learning and hybrid model development. Previous research, such as the work by Bakhtiari et al., has reviewed and compared various kinetic expressions involving different dependent variables concentrations (C_i), fugacities (f_i), or partial pressures (P_i). For each model, the corresponding equilibrium constant (K_{eq}) must be specified.

Given the reversibility of the methanol dehydration reaction, selecting an appropriate kinetic model and developing a robust mathematical framework are essential for accurate equilibrium predictions, as these are strongly influenced by how the equilibrium constant is calculated. Once Keq is determined, the equilibrium conversion (X_{eq}) can be calculated using the inverse relation. In addition to kinetic and equilibrium expressions, the modeling framework also requires definitions of thermophysical properties such as specific heat capacity, thermal conductivity, and viscosity for both individual components and the overall mixture, as well as heat transfer equations. It is also important to highlight that, in the case of the water-cooled reactor configuration, the coolant temperature on the shell side remains constant. The use of saturated water ensures steady heat removal,

though it may lead to the formation of a two-phase system due to boiling, which must be considered when designing the heat exchange process.

A single unified model was trained to predict all output variables simultaneously, including outlet concentrations of methanol, dimethyl ether, and water, reactor temperature, pressure, and overall conversion. This integrated approach allows the model to capture interdependencies and physical correlations among the process variables, enhancing prediction consistency and interpretability within the hybrid and data-driven modeling frameworks.

Prior to generating the synthetic dataset, the developed first-principles model was rigorously validated against experimental data reported in the literature for DME synthesis over γ-Al_2O_3 in a fixed-bed reactor [59]. the model accurately reproduces the experimental temperature and conversion profiles under identical operating conditions (T_0 = 550 K, P_0 = 2.1 bar, F_{T_0} = 0.145 kmol/h, yM_0 = 1), confirming its physical fidelity and reliability for subsequent data generation and hybrid modeling.

Mass Balance: $$-\frac{1}{A_C}\frac{dF_i}{dz} + \rho_b R_i = 0 \qquad i = \text{MeOH, DME, Water} \tag{19}$$

Energy Balance: $$-\frac{C_p}{A_c}\frac{d(F_t T)}{dz} + \rho_b \sum_{i=1}^{3} R_i\,(-\Delta H_{f,i}) = 0 \tag{20}$$

Momentum Equation: $$\frac{dP}{dz} = \frac{150\mu}{\varnothing_s^2 dp^2}\frac{(1-\varepsilon)^2}{\varepsilon^3}u + \frac{1.75\rho}{\varnothing_s dp}\frac{(1-\varepsilon)}{\varepsilon^3}u^2 \tag{21}$$

Initial Conditions (Reactor Entrance): $$F_i = F_{i,0};\quad F_t = F_{t,0};\quad T = T_0;\quad P = P_0 \tag{22}$$

$$F_T = \sum_{i=1}^{3} F_i \quad \rightarrow dF_T = d\sum_{i=1}^{3} F_i \tag{23}$$

$$КРЕП = k_s f_M^{0.55} \left[1 - \frac{f_D f_W}{K_{eq} f_M^2}\right] \quad \frac{mol}{g_{catalyst}.h} \tag{24}$$

$$k_s = 1457.024 \exp\left(-\frac{78072.55}{RT}\right) \quad \frac{mol}{g_{catalyst}.h.Pa^{0.55}} \tag{25}$$

$$\ln\left(K_{eq}\right) = \frac{2835.2}{T} + 1.675 \ln(T) - 2.39 \times 10^{-4} - 0.21 \times 10^{-6} T^2 - 13.3 \tag{26}$$

Table.2. Constants and initial values for the reaction in dimethyl ether fixed-bed reactor

Variable	Unit		Variable	Unit	
Temperature	K	551.15	Reactor diameter	m	0.078
Pressure	bar	2.1	Catalyst diameter	m	0.003
Total volume flow rate	dm^3/h	4.34	Bed porosity	-	0.4
Initial mole fraction of methanol	-	1	Catalyst density	kg/m^3	1470
Reactor length	m	0.7	Bed density	kg/m^3	882

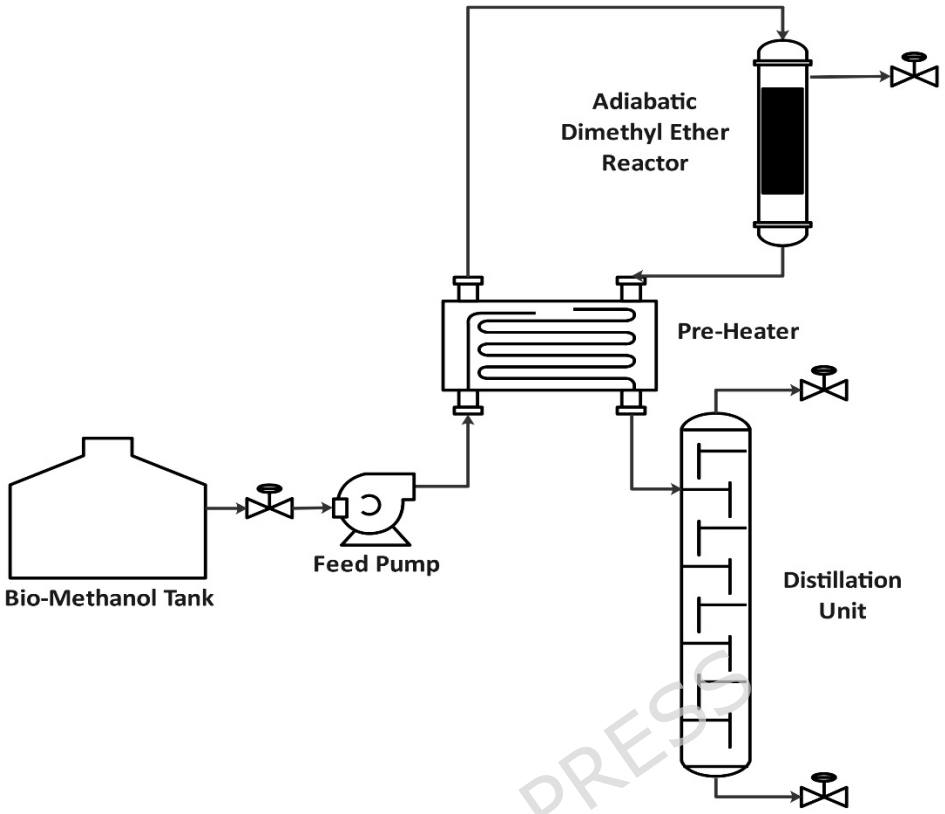


Fig.4. The schematics of adiabatic dimethyl ether systems

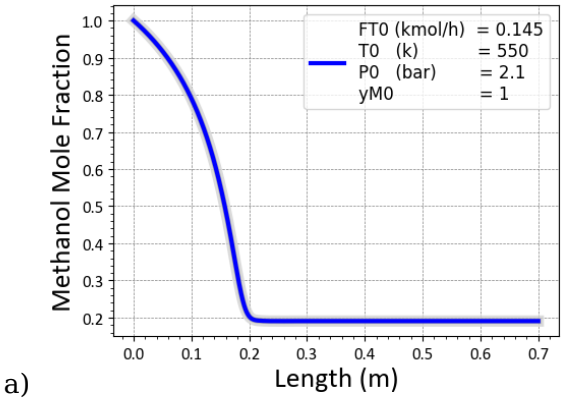
3. Results and Discussion

3.1. First-Principles Modeling

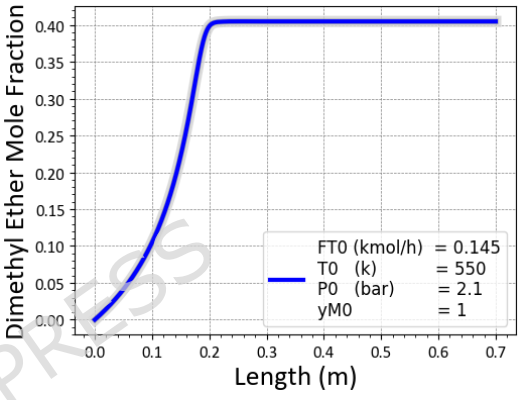
The first-principles results were obtained by solving the system of differential equations described in Section 2.6, using thermophysical properties from supplementary tables and operating conditions sampled via Latin Hypercube Sampling (LHS; Section 2.5). As previously validated against experimental data (Fig. 6), the model was used to generate 7,000 synthetic data points under varied non-ideal conditions.

As shown in Fig. 5, the reactor exhibits expected trends along its 0.7 m length: conversion rate, product concentrations (water and dimethyl ether), and temperature increase, while pressure and methanol concentration

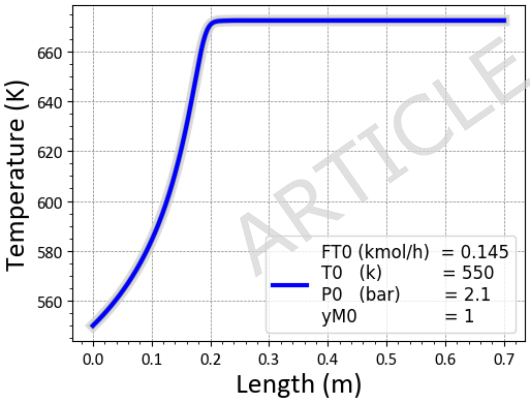
decrease. Due to the short reactor length and low flow rate, pressure drop is minimal (~0.005 bar). The final conversion reaches 80%, with water and dimethyl ether mole fractions of 0.4 each, and methanol reduced to 0.2. The outlet temperature is 673 K, consistent with the exothermic reaction profile. These results confirm the model's ability to capture realistic reactor behavior under non-ideal conditions.



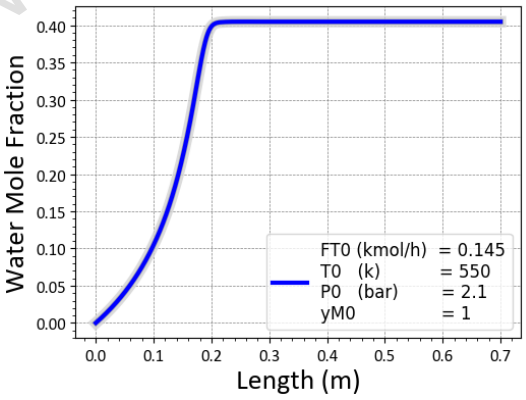
a)



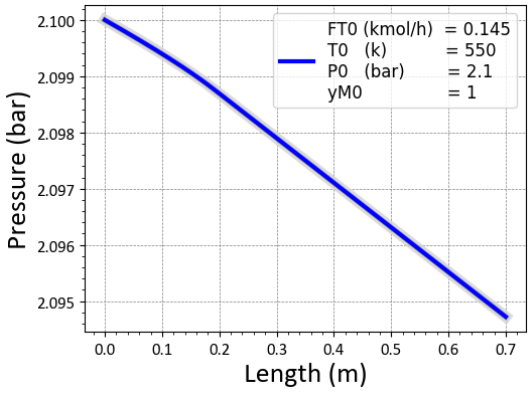
b)



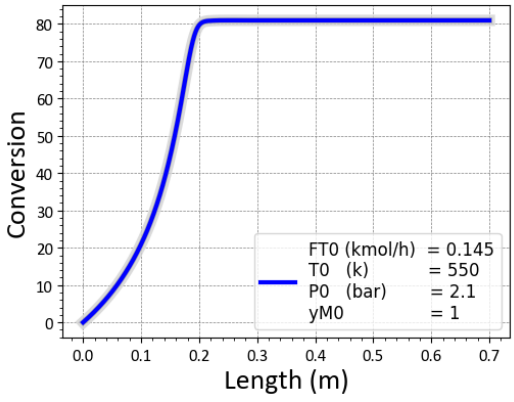
)d



c)



e)



f)

Fig.5. Fixed-bed dimethyl ether reactor (L = 0.7 m) outputs using the first principles model by basic initial conditions (T_0 = 550 K, P_0 = 2.1 bar, F_{T_0} = 0.145 kmol/h, y_{M_0} = 1); methanol mole fraction (a), dimethyl ether (b), water (c), temperature (d), pressure (e), and conversion % (f)

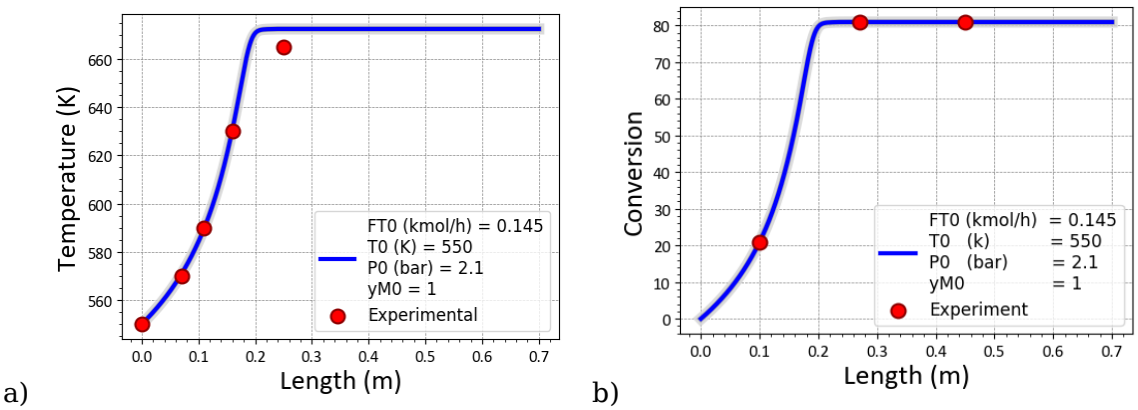


Fig. 6. Validation of the first-principles model against experimental data [59] for DME fixed-bed reactor. (a) Temperature and (b) conversion profiles along the reactor length (L = 0.7 m); initial conditions ((T_0 = 550 K, P_0 = 2.1 bar, F_{T_0} = 0.145 kmol/h, y_{M_0} = 1)

3.2. Hybrid Modeling

3.2.1 Hybrid Modeling (Correction; Extrapolation Feature)

Since the data within the reactor process are sequential and have a spatial order (the reactor output at each location depends on the outputs at previous locations), this type of network can easily model the process. The LSTM network in the correction model functions as a hybrid correction–extrapolation mechanism, receiving the outputs of the first-principles model (e.g., conversion and temperature profiles along the reactor length) along with relevant operating parameters, and learning the residual mapping between mechanistic predictions and observed (or synthetic) data, rather than operating as an independent surrogate. Here, using a long short-term memory (LSTM) neural network, which is a type of RNN with long-term memory, we modeled the dimethyl ether production process under one set of operational conditions. The hybrid models were trained on the 7,000

synthetic data points (65% train, 10% validation, 25% test; Section 2.5) generated from the validated first-principles simulator (Fig. 6). We observed that this network could accurately predict the process outputs at subsequent locations. Essentially, this network can perform extrapolation effectively, using spatial/temporal or sequential data to predict future outputs. As shown in Fig. 7, the LSTM network accurately predicted five different patterns in the process outputs, including temperature, conversion rate, and three reactant and product components. With an MSE error parameter of 45e-5, the network can perform the necessary extrapolation to predict the reaction outputs at subsequent locations. Given the explanations about this neural network, it uses its short-term memory and previous outputs to perform extrapolation with high accuracy, aiming to predict the expected future outputs. Table 3 lists the input and output variables of this network along with their variation ranges. The accuracy and error rate comparison of the model were conducted with the first-principles model simulation, and as expected, the hybrid model can easily replace the aforementioned first-principles model.

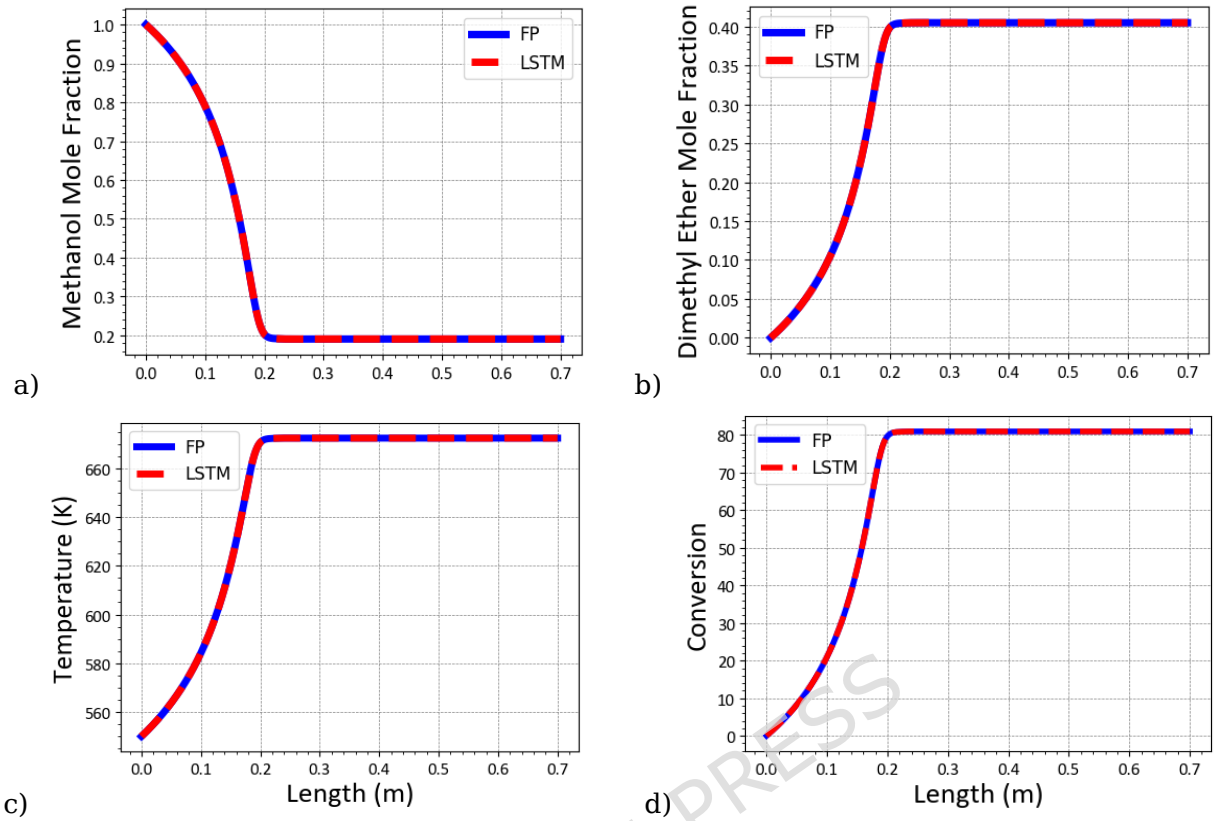


Fig.7. Comparison of fixed-bed dimethyl ether reactor output using correction (LSTM) and first principles model by basic initial conditions; methanol mole fraction (a), dimethyl ether (b), temperature (c), and conversion (d)

Table.3. LSTM model input / output of fixed-bed dimethyl ether reactor with range of variations

	Variable	Symbol	Unit	Range
Inputs	Catalyst bed length	z	m	0 – 0.7
	Temperature (z)	T	K	550 – 690
	Mole fraction of methanol (z)	y_M	-	0.2 – 1
	Mole fraction of water (z)	y_W	-	0 – 0.4
	Mole fraction of DME (z)	y_{DME}	-	0 – 0.4
	Conversion (z)	X	-	0 – 80
Outputs	Temperature (z+dz)	T	K	550 – 690
	Mole fraction of methanol (z+dz)	y_M	-	0.2 – 1

Mole fraction of water (z+dz)	y_W	-	0 – 0.4
Mole fraction of DME (z+dz)	y_{DME}	-	0 – 0.4
Conversion % (z+dz)	X	-	0 – 80

3.2.2 Hybrid Modeling (Estimation; Rate Reaction Estimation Feature)

In this section, we discuss another type of hybrid model, namely the estimation of unknown models or parameters. Most reaction kinetics relationships are derived from experiments and are largely empirical. Even in an industrial setting, an experimental model can be implemented using experimental or industrial data. In the dimethyl ether production reactor studied here, the provided reaction kinetics equations are empirical. By using rate constants, calculating the fugacity of each component at each location (in the unsteady-state reactor at each time), and applying the given relationship, the spatial reaction rate can be determined.

Due to non-ideal conditions, the Peng-Robinson equation of state is used to calculate material properties at any location and time. However, several fundamental challenges arise: non-ideal conditions requiring the use of an equation of state for physical and chemical property calculations, reliance on an empirical rate relationship, and the additional complexity of property calculations in an unsteady-state reactor.

To address these challenges, we leverage data generated from first-principles modeling of the reactor and select relevant inputs along with their variation ranges to estimate the reaction rate at each location. We implement a model using three different machine learning algorithms (XGB, KNN, GBR) to predict the spatial reaction rate based on specified process inputs. This hybrid approach enables us to replace the empirical reaction rate relationship with a data-driven model. At each point in the reactor, the output of the machine learning model (the reaction rate) is first obtained and then

integrated into the governing equations, which enter the set of differential equations to compute the reactor's output at each location.

If experimental data for reaction kinetics are available, this hybrid model can be used to efficiently calculate the reaction rate at any moment and location without relying on complex kinetic relationships. Instead of depending on empirical kinetic models that may have inherent limitations, our approach directly predicts the reaction rate using machine learning, trained on available process data. This not only enhances prediction accuracy but also facilitates process optimization by eliminating the need for deriving intricate kinetic relationships and conducting extensive experiments to estimate kinetic parameters.

As shown in Table 6, the proposed hybrid model achieves high accuracy, with minimal error when compared to the first-principles model. This allows the machine learning model to serve as a reliable alternative for reaction kinetics estimation in various applications, including process optimization. Furthermore, as illustrated in Fig. 7, the first-principles and hybrid models exhibit strong agreement across the six considered outputs, confirming the hybrid model's ability to capture the process behavior with high precision.

Table.4. Estimation model (data-driven section) output with R^2 and MSE for fixed-bed dimethyl ether reactor

Output Variable	Symbol	Unit	MSE (XGB)	R^2 (XGB)	MSE (KNN)	R^2 (KNN)	MSE (GBR)	R^2 (GBR)
Reaction rate	r_i	$mol/g_{catalyst}.h$	12.54e-8	0.999	0.53e-8	0.999	374e-8	0.999

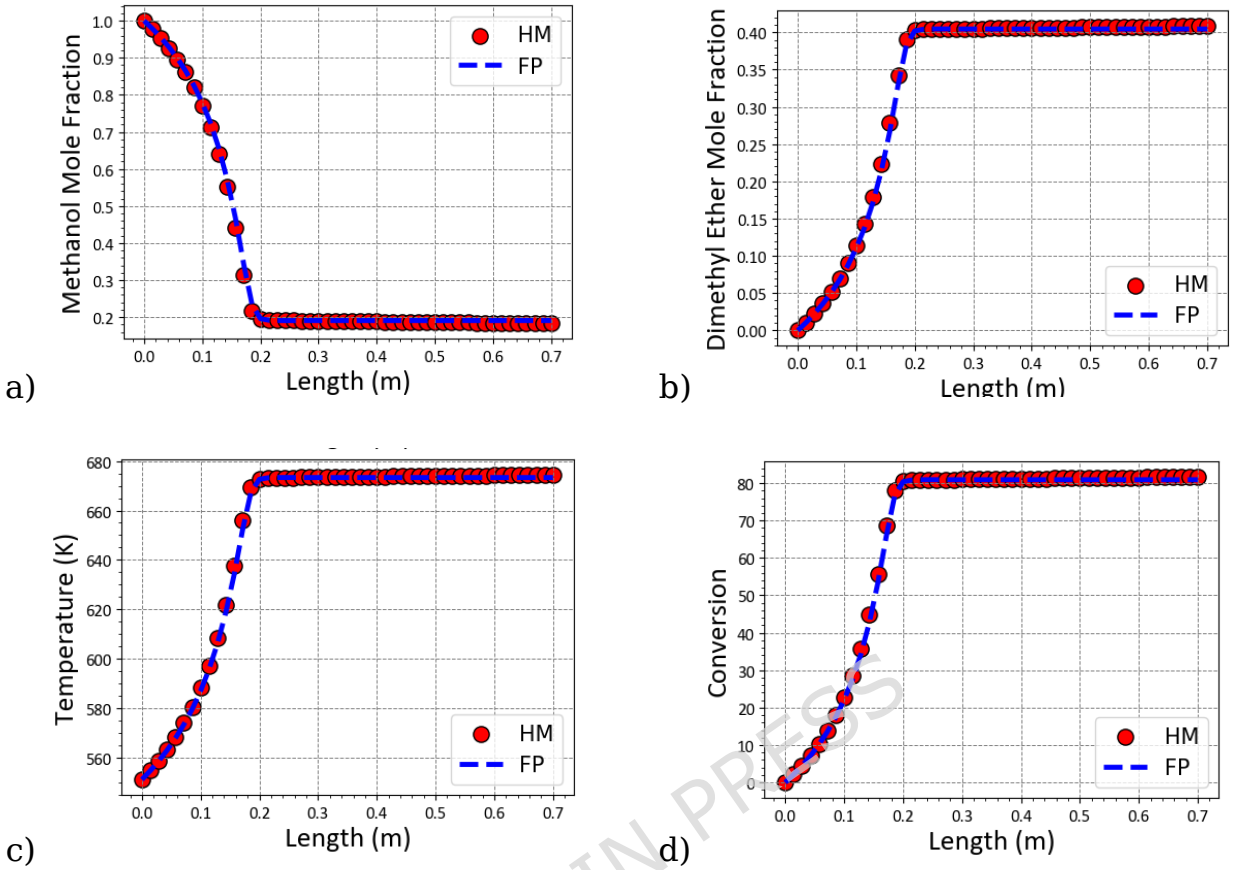


Fig.8. Comparison of fixed-bed dimethyl ether reactor output using estimation and first principles model by basic initial conditions; methanol mole fraction (a), dimethyl ether (b), temperature (c), and conversion (d)

Table.5. Estimation model input / output (data-driven section) of fixed-bed dimethyl ether reactor with range of variations

Output Variable	Symbol	Unit	Range
Mole fraction of methanol	y_M	-	0.2 – 1
Mole fraction of water	y_W	-	0 – 0.4
Mole fraction of DME	y_{DME}	-	0 – 0.4
Temperature	T	K	550 – 670
Pressure	P	bar	2.09 – 2.1
Reaction Rate	r_i	$mol/g_{catalyst}.h$	-9.1 – 0.54

Table.6. Estimation model output with R^2 and MSE for fixed-bed dimethyl ether reactor

Output Variable	Symbol	Unit	R^2	MSE
Mole fraction of methanol	y_M	-	0.999	659e-8
Mole fraction of water	y_W	-	0.999	164e-8
Mole fraction of DME	y_{DME}	-	0.999	164e-8
Temperature	T	K	0.999	1.395
Pressure	P	bar	0.999	4e-11
Conversion	X	-	0.999	0.659

3.3. Data-Driven Modeling

To demonstrate that black-box models can effectively replace first-principles models, we proceed with the first hybrid modeling approach known as surrogate modeling. Here, we first select the inputs and outputs of the surrogate model using a case study and sensitivity analysis. Using the first-principles model presented in the previous section, we generate data and then use three machine learning algorithms (KNN, XGB, and GBR) to create the surrogate model for this process. The superior performance of KNN, GBR, and XGB (Table 7) validates their selection for steady-state surrogate modeling, as justified in Section 2.2. This models (KNN, GBR, XGB) were applied to the same dataset (Section 2.5) to replace the full mechanistic solver, with performance reported in Table 7. These three algorithms were chosen due to their higher accuracy and greater efficiency compared to other available machine learning algorithms. As mentioned, for data-driven modeling and surrogate modeling, three different algorithms (XGB, KNN, and GBR) were used. The results, including the error rates between the experimental data and the predictions made by the models, are detailed in Table.7 using two error measurement parameters (MSE, R^2). This table

clearly shows that all three models have very high accuracy for all six outputs, with no significant differences in the outputs. It should be noted that the accuracy of the KNN model is approximately 30 times higher than that of the XGB model and twice as high as that of the GBR model. However, since the overall error for all three models is negligible, the other two models can also be used. To better understand the error rates of each model, we plotted the predicted data against the experimental data. It is important to note that these surrogate models can easily replace the main model. Given that the selected input and output variables for the process consider different operational conditions such as varying temperatures and pressures, bed porosity, reactor diameter, and different molar fractions of water and methanol, this model can be used for predictions and optimizations under various operational conditions. Given the very high accuracy of these three models, their use will practically involve no error, yielding desirable results.

Table.7. Performance comparison of Data-Driven models with R^2 and MSE for fixed-bed dimethyl ether reactor

			Test		**Train**		**Test**		**Train**		**Test**		**Train**	
Output Variable	Symbol	Unit	MSE (XGB)	R^2 (XGB)	MSE (XGB)	R^2 (XGB)	MSE (KNN)	R^2 (KNN)	MSE (KNN)	R^2 (KNN)	MSE (GBR)	R^2 (GBR)	MSE (GBR)	R^2 (GBR)
Mole fraction of methanol	y_M	-	35e-8	0.999	15e-8	0.999	18e-8	0.999	50e-8	0.999	25e-8	0.999	10e-8	0.999
Mole fraction of water	y_W	-	79e-8	0.999	52e-8	0.999	45e-8	0.999	20e-8	0.999	69e-8	0.999	20e-8	0.999
Mole fraction of DME	y_{DME}	-	13e-8	0.999	75e-8	0.999	45e-8	0.999	20e-8	0.999	89e-8	0.999	10e-8	0.999
Temperature	T	K	0.844	0.999	0.604	0.999	0.05	0.999	0.04	0.999	0.0771	0.999	0.01	0.999
Pressure	P	bar	6e-6	0.999	3e-6	0.999	1e-6	0.999	5e-6	0.999	10e-6	0.999	1e-6	0.999
Conversion	X	-	35e-6	0.999	25e-6	0.999	18e-6	0.999	1e-6	0.999	25e-6	0.999	1e-6	0.999

Table.8 Data-Driven models input / output of fixed-bed dimethyl ether reactor with range of variations

	Variable	Symbol	Unit	Range
Inputs	Catalyst bed length	z	m	0 – 0.7
	Total molar flow rate	F_{T0}	kmol/h	0.1 – 0.19
	Initial temperature	T_0	K	540 – 620
	Initial pressure	P_0	bar	0.7 – 2.1
	Bed porosity	ε_b	-	0.2 – 0.8
	Diameter	D	m	0.07 – 0.15
	Initial mole fraction of methanol	y_{M0}	-	0.3 – 1
	Initial mole fraction of water	y_{W0}	-	0 – 0.25
Outputs	Mole fraction of methanol	y_M	-	0.1 – 0.98
	Mole fraction of water	y_W	-	0 – 0.7
	Mole fraction of DME	y_{DME}	-	0 – 0.4
	Temperature	T	K	520 – 720
	Pressure	P	bar	1.9 – 2.5
	Conversion	X	-	0 – 87

3.4. Data Quantity on Model Accuracy

Aside from examining various sampling methods and their impact on the results of data-driven models, another factor that needs to be evaluated is the number of data points available in the dataset. Generally, to evaluate this influencing parameter using two error calculation metrics mean squared error (MSE) and coefficient of determination (R^2) and the XGB machine learning model implemented for the dimethyl ether production reactor modeling, we conducted this task. We used a step size of 500 data points and a dataset of 70,000 data points generated by the first-principles model of the dimethyl ether production reactor under the initial conditions mentioned in the results and discussion section of the model. Each time, using a new

dataset (starting with 500 data points and adding more data in each iteration by the step size), we implemented the XGB machine learning model. As shown in Fig.9, from the dataset with 5000 data points onward, our model exhibited low error and high accuracy, and from 10,000 data points until the end, the accuracy and error rate of our model remained unchanged. Two points are crucial here: to enhance the accuracy and maintain randomness when adding new data to the dataset, we used the random shuffle function from the NumPy library, which shuffles the data each time before adding them to the dataset. Notably, both charts exhibit slight fluctuations due to rounding errors in the programming language (due to the small values) and the diversity of data in the dataset, although the overall trend shows a decrease in error.

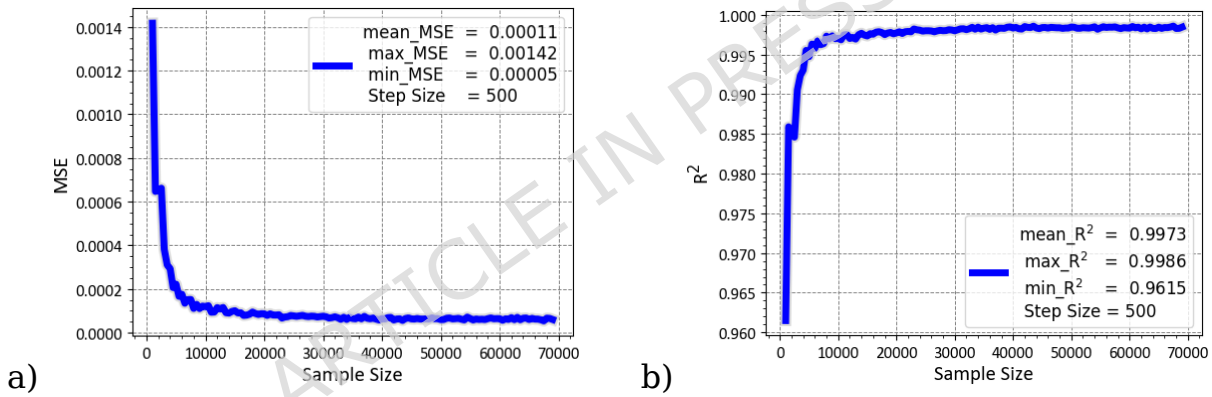


Fig.9. Variations of the MSE (a) and the R^2 (b) to amount of data in the data-driven model

3.5. Sensitivity Analysis

The Pearson correlation coefficient is a parametric statistical method that indicates the strength and direction of the relationship between two variables. Like other correlation methods, this method considers the relationships between pairs of variables, evaluating the relationship between two variables A and B with or without the presence of a third variable like C, yielding a consistent value. This coefficient measures the relative correlation

between two variables, ranging from -1 to +1. If the value obtained is positive, changes in the two variables occur in the same direction; in other words, as one variable increases, the other also increases. Conversely, if the value is negative, the two variables act in opposite directions. It is worth noting that a value of zero indicates no relationship between the two variables, while a value of +1 indicates a perfect positive correlation and -1 a perfect negative correlation. [72, 73]

As observed in the case study section, we use the Pearson correlation coefficient to examine the correlation between each process input variable and its six outputs. Fig.10 clearly shows the correlation coefficient for all variables. As expected, the input variables such as reactor length, molar fraction of methanol and water, and feed inlet temperature have a more significant impact on the process outputs compared to other variables. The closer the correlation coefficient value is to +1 or -1, the greater the impact. For example, for feed inlet temperature, as the temperature increases, the reaction rate and output increase accordingly, reflected by the positive correlation coefficients for the molar fractions of water, dimethyl ether, and conversion rate (0.18, 0.4, 0.36).

$$r(X_i,Y) = \frac{\sum_{j=1}^{n}(X_{i,j} - \bar{X}_i)(Y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{n}\left(X_{i,j} - \bar{X}_i\right)^2 \sum_{j=1}^{n}\left(Y_j - \bar{Y}\right)^2}} \tag{27}$$

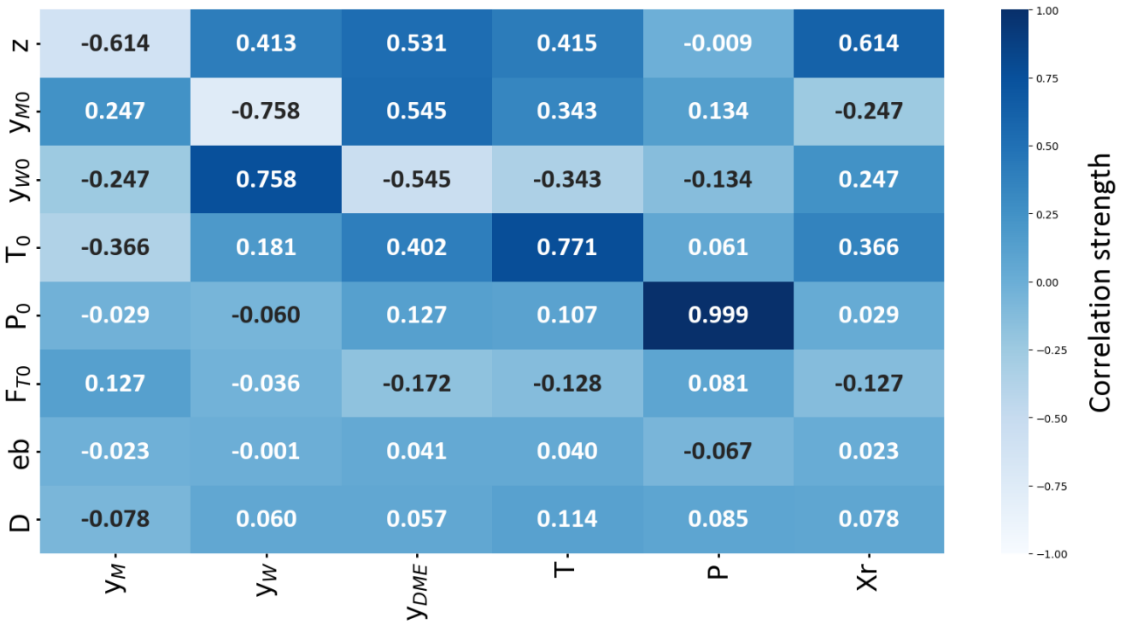


Fig.10. Sensitivity analysis of the input / output variables for data-driven model

3.6. Case Study of Process Variables

In the subsequent discussion, after first-principles modeling of the process, to generate data for use in black-box (surrogate) modeling and hybrid parameter estimation modeling, we first conducted a case study on all process variables using the data-driven model. As presented in Table 9, which outlines the input variables along with the range of output variations, our proposed data-driven model effectively predicts the process output in response to input variations, similar to the first-principles model. This demonstrates that the data-driven approach can serve as a highly suitable alternative, allowing us to incorporate both operational and process variables into process modeling.

Generally, increasing the reaction progress rate will result in reaching the desired temperature, pressure, molar fractions of reactants and products, and final conversion rate over a shorter reactor length. Conversely, decreasing the reaction progress rate will require a longer reactor length to achieve the final variable values. increasing the molar feed rate into the

reactor causes the desired conversion rate to be reached over a longer reactor length. Additionally, as we know, an increased feed rate leads to higher velocity and consequently a higher pressure drops along the reactor. Therefore, the total molar feed rate will affect the reaction outputs. the effect of reactor diameter on the output. As we know, a larger reactor diameter results in a lower pressure drop and faster reaction, leading to a quicker achievement of the final conversion rate.

The bed porosity also has a direct relationship with the reaction rate; as bed porosity increases, the likelihood of catalyst particle contacts and the required surface area decrease, resulting in the reaction achieving the final conversion rate over a longer reactor length. Increasing the initial molar fractions of all three components (methanol, water, and dimethyl ether) reduces the reaction rate for producing dimethyl ether. If the molar fraction of any component increases at the reactor inlet and the start of the reaction, the desired conversion rate will be achieved over a longer reactor length. It is important to note that changes in the inlet water molar fraction have a negligible effect on the process output, and decreasing the inlet methanol molar fraction will increase the overall reaction conversion rate due to the reduced total methanol amount while maintaining the reaction rate. Finally, increasing the inlet temperature and pressure enhances particle collisions and thus increases the reaction rate. This increased rate helps achieve the final conversion rate and other output variables over a shorter reactor length. Finally, in Figures 11 to 13, it is clearly observed how variations in three variables total initial molar flow rate, reactor diameter, and bed porosity affect the process outputs. The impact of their increase or decrease on process performance is also well illustrated.

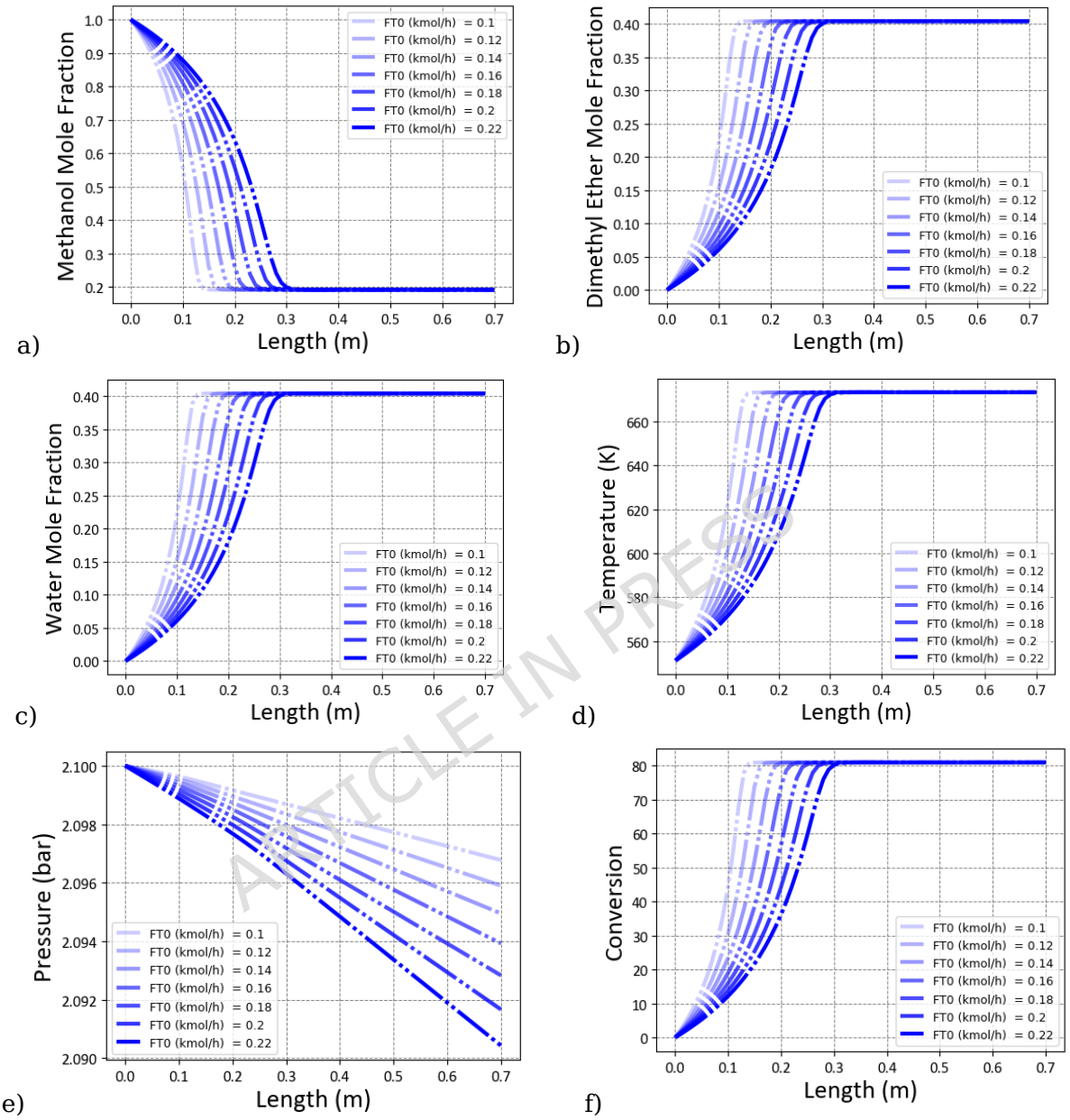


Fig.11. Reactor outputs for the case study on the variable initial total molar flow rate for the dimethyl ether fixed-bed reactor; molar fraction of methanol (a), dimethyl ether (b), water (c), temperature (d), pressure (e), and conversion (f).

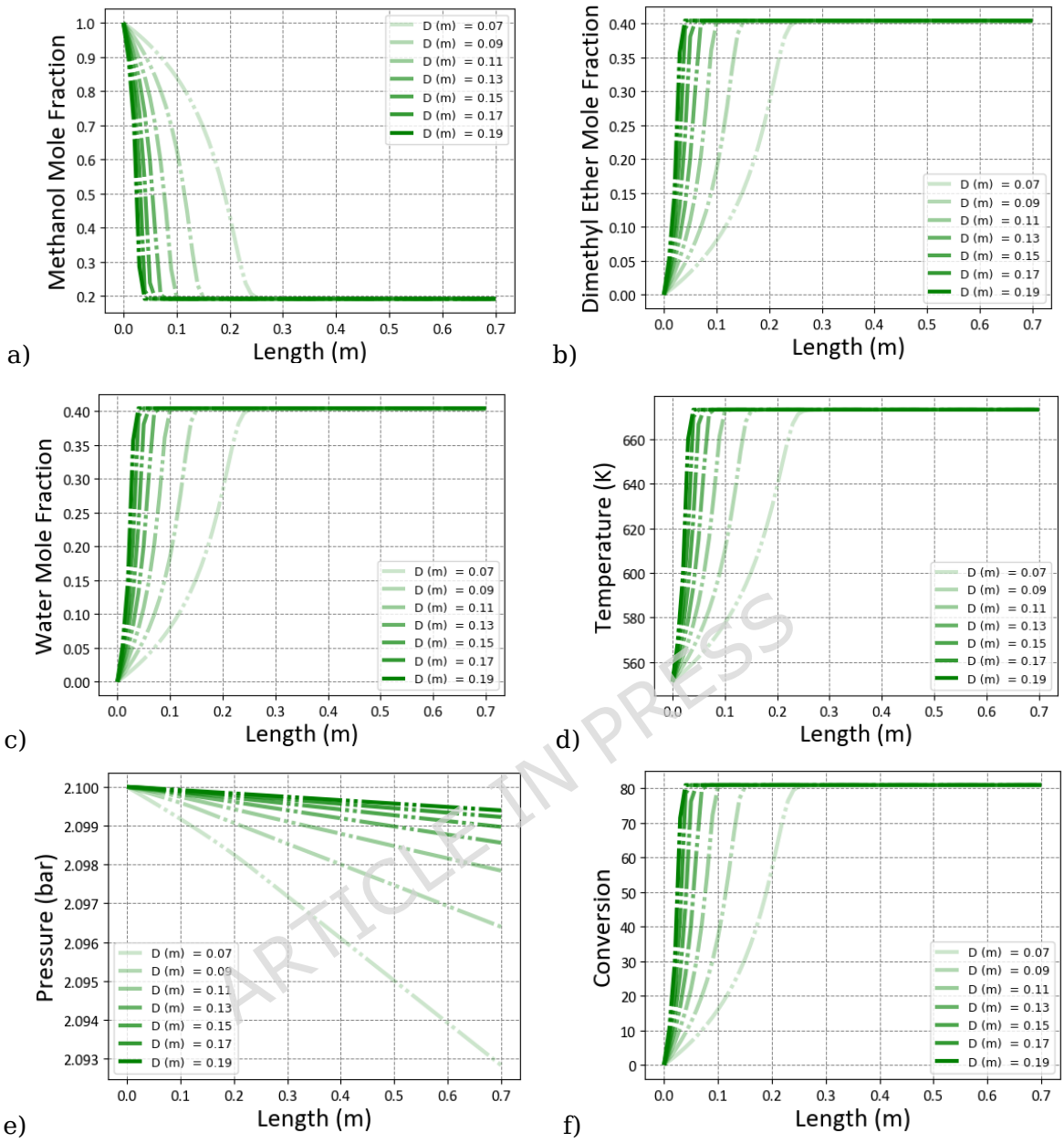


Fig.12. Reactor outputs for the case study on the variable reactor diameter for the dimethyl ether fixed-bed reactor; molar fraction of methanol (a), dimethyl ether (b), water (c), temperature (d), pressure (e), and conversion (f).

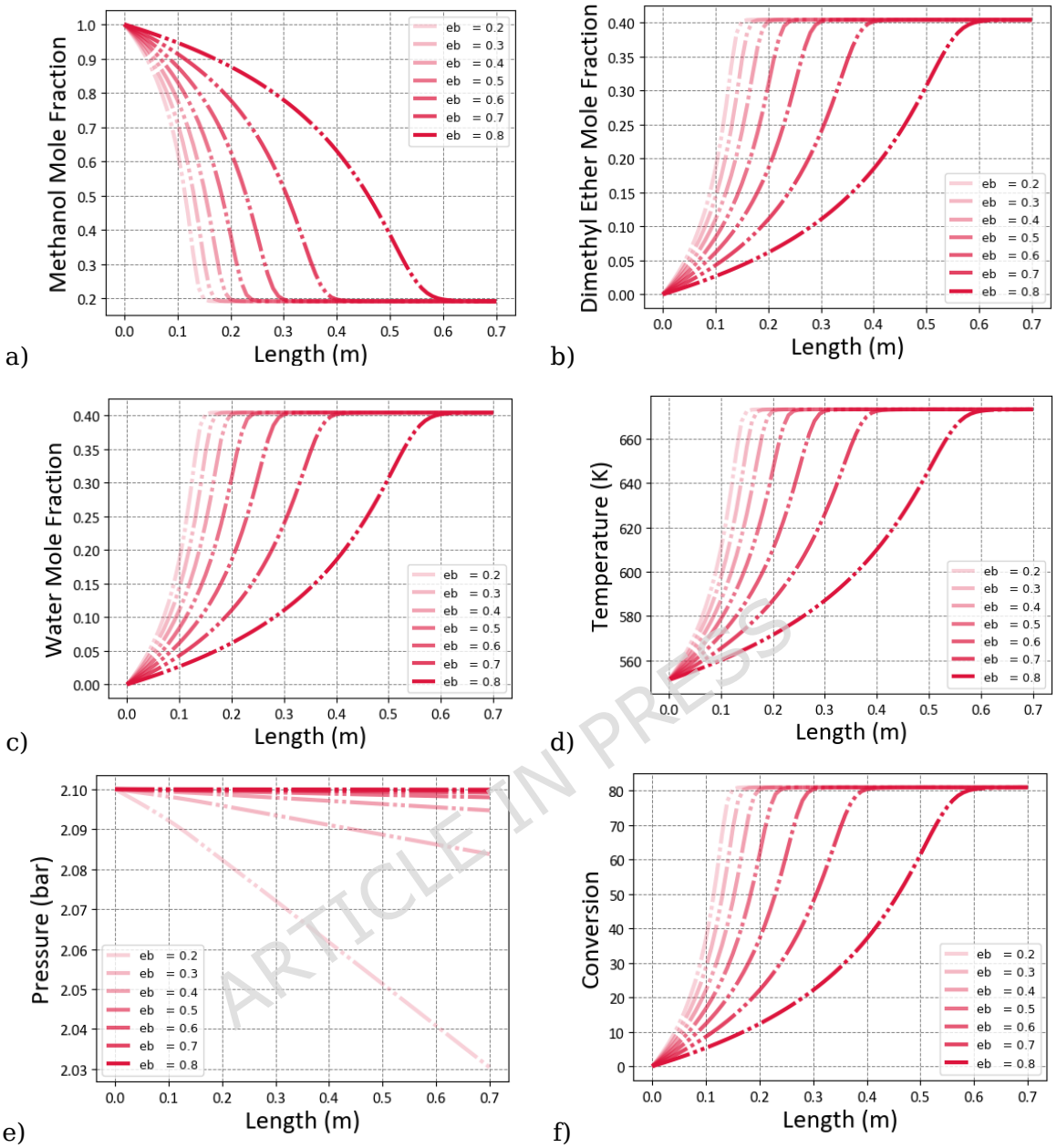


Fig.13. Reactor outputs for the case study on the variable bed porosity for the dimethyl ether fixed-bed reactor; molar fraction of methanol (a), dimethyl ether (b), water (c), temperature (d), pressure (e), and conversion (f).

Table.9. Variables used in the case study of fixed-bed dimethyl ether reactor with range of variations

Variable	Symbol	Unit	Range
Total molar flow rate	F_{T0}	kmol/h	0.1 – 0.19
Initial temperature	T_0	K	540 – 620
Initial pressure	P_0	bar	0.7 – 2.1
Bed porosity	ε_b	-	0.2 – 0.8
Reactor diameter	D	m	0.07 – 0.15
Initial mole fraction of methanol	y_{M0}	-	0.3 – 1
Initial mole fraction of water	y_{W0}	-	0 – 0.25
Initial mole fraction of DME	y_{DME0}	-	0 – 0.6

3.7. Optimization

Because data-driven models can be used at different operation conditions and less run-time, these models, unlike complex mechanistic models, have been widely used for process optimization, after reviewing and comparing performance different data-driven models for unseen data, the XGB model error was less and also the investigation of various optimization methods around different processes in chemical engineering [74, 75], by the objective function of the Maximum Conversion at Minimum Temperature (MC-MT) which includes the function of T, X ($f_{(T, X)}$), and with a suitable initial guess and using different constraint for each output species, first by Differential Evolution (DE) method [74, 75] and then to compare the optimal values with 2 other methods by SLSQP [76] and TRUST-CONSTR [77] where each initial guess was used from the answer of DE method and implemented and optimized using the functions available in SciPy package and the output values (Table 10) and optimal input (Table 11) for each of the variables are reported. As can be seen from the optimization results, in the implemented methods, the DE method provided the most optimal solution for us, both in

terms of (MC-MT) and conversion% and according to the amount of MC-MT in all methods, the amount 0.1534 will be the optimal value for 84.3 conversion%.

$$\text{MC - MT: } f_{(T,\,X)} = \frac{T_{out}}{T_{Nominal}} - \frac{X_{out}}{X_{Nominal}} \tag{28}$$

Our objective in optimizing this process is twofold: first, to demonstrate the application of data-driven models for faster optimization that includes the process's operational conditions, and second, to obtain the optimal process input variable values, such as initial temperature and pressure, methanol molar fraction, reactor diameter and bed porosity, as well as total molar flow rate and so on. As expected, by performing this operation, we could calculate the optimal input variable values of (521 K, 2.1 bar, 0.825, 0.139 m, 0.593 and 0.172 kmol/h) for the lowest output temperature (605.9 K) and the highest conversion (84.3%). As can be seen in Table 10, under the optimal process conditions, we achieved the maximum conversion at the lowest temperature. Additionally, by increasing the total molar flow rate (0.172 kmol/h), we can produce the maximum amount of dimethyl ether. Furthermore, by setting another objective function for optimization, we can achieve desired values under different operational conditions of the process and scale-up it.

To visualize the convergence behavior of the Differential Evolution (DE) optimization algorithm, the scatter convergence profile is presented in Figure 14. This figure depicts the variation of the objective function value across successive iterations, demonstrating the algorithm’s stable convergence toward the global optimum and confirming the robustness of the optimization procedure.

Table 10. Optimal output values for each optimization method

Name	Unit	DE	TRUST-CONSTR	SLSQP	Nominal Operation

Mole fraction of methanol	-	0.21	0.137	0.13	0.2
Mole fraction of water	-	0.395	0.704	0.72	0.4
Mole fraction of DME	-	0.385	0.153	0.15	0.4
Pressure	bar	2.1	2.096	2.09	2.1
Temperature	K	605.9	622	620	673
Conversion	-	84.3	86.4	86	80
MC-MT	-	0.1534	0.1557	0.1537	

Table 11. Optimal input values for each optimization method

Name	Unit	DE	TRUST-CONSTR	SLSQP	Nominal Operation
Catalyst bed length	m	0.7	0.7	0.7	0.7
Total molar flow rate	kmol/h	0.172	0.206	0.2	0.145
Initial temperature	K	521	551	551	550
Initial pressure	bar	2.1	2.1	2.1	2.1
Bed porosity	-	0.593	0.5	0.5	0.4
Diameter	m	0.139	0.1	0.1	0.078
Initial mole fraction of methanol	-	0.825	0.44	0.4	1

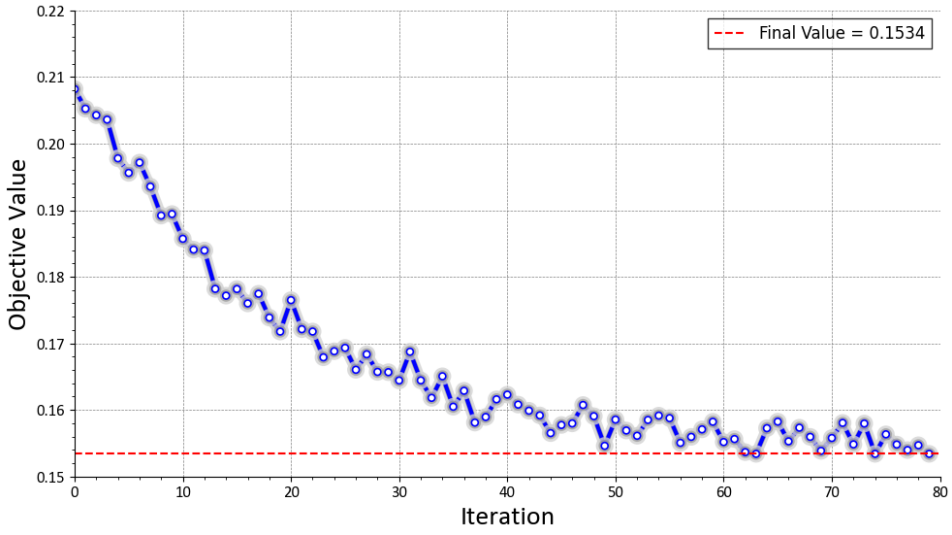


Fig.14. Scatter convergence profile of the DE optimization algorithm showing the variation of the objective function value with iteration number and its stable convergence toward the optimal solution.

3.8. Computational Efficiency Analysis

One of the major advantages of applying data-driven and hybrid modeling frameworks in chemical process simulation is their computational efficiency compared with conventional mechanistic solvers. To quantitatively evaluate this benefit, timing benchmarks were conducted for the first-principles simulator and each developed model under identical input conditions.

The mechanistic reactor model, implemented using the solve_ivp function with the BDF integration method, required an average runtime of 0.9253 seconds to compute the full reactor profile containing 701 axial discretization points. In contrast, the trained data-driven and hybrid models achieved inference times of less than 0.06 seconds per simulation, providing up to a 24× speedup while maintaining predictive accuracy.

The comparative results are summarized in Table 12, which reports the average runtimes from 100 repeated evaluations for statistical consistency.

Table 12. Computational performance comparison between the first-principles, data-driven, and hybrid models.

Model	Type	Avg. Runtime (s)	Speedup vs. Mechanistic
Governing Equation	First-Principles	0.9253	×1
XGB	Data-Driven (Inference)	0.042	~×22
KNN	Data-Driven	0.0381	~×24
GBR	Data-Driven	0.0456	~×20
Estimation	Hybrid	0.0489	~×19
Correction	Hybrid	0.0437	~×21

The results clearly demonstrate that all data-driven and hybrid models achieve ×18–24 faster runtime performance relative to the first-principles model. Such computational efficiency makes these models highly suitable for iterative optimization, real-time process control, and large-scale uncertainty analysis applications where direct numerical integration of governing equations would be computationally prohibitive.

It should be noted that the comparison excludes one-time model training costs for machine learning algorithms, as runtime efficiency during inference is the dominant factor for operational use and optimization-based deployment. The significant speed improvement validates the practical utility of hybrid and data-driven frameworks for future digital twin and real-time optimization applications in chemical reaction engineering.

3.9. SHAP Analysis

To further understand the influence of each input feature on the model outputs, SHAP (**SH**apley **A**dditive ex**P**lanations) [78] values were computed. SHAP values provide both the direction and magnitude of a feature's impact on the model prediction for each instance. In Figure 15, SHAP summary plots are presented for six target variables, where each point represents an individual sample, colored by the feature value (from low: blue to high: red). The horizontal axis indicates the SHAP value, reflecting the effect of that feature on the model output.

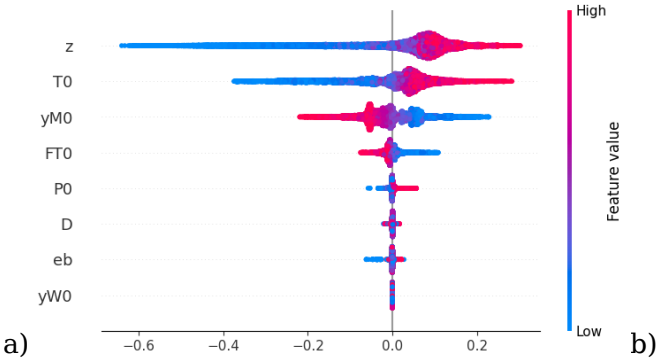
In plot (a), the SHAP analysis for the conversion (Xr) clearly shows that the reactor length (z), initial temperature (T_0), and initial molar fraction of methanol (y_{M0}) are the most influential variables. High values of Catalyst bed length (z) and initial temperature (T_0) lead to positive SHAP values, implying that increasing these variables enhances the conversion rate. This finding aligns with chemical engineering principles: a longer reactor provides more

residence time, facilitating reaction completion, while higher inlet temperature increases reaction kinetics.

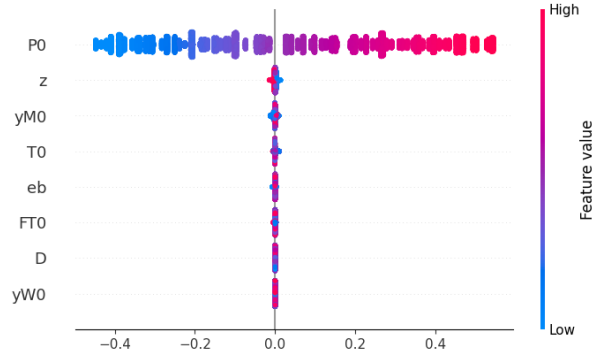
Interestingly, we observe a dense cluster of blue points on the left-hand side of the z feature, indicating that lower values of z consistently result in negative SHAP values, thus reducing conversion. This strong sensitivity to reactor length highlights its pivotal role in system performance. Similarly, higher values of y_{M0} (red points) also correspond to positive SHAP values, confirming that greater methanol content in the feed boosts conversion efficiency consistent with the reaction stoichiometry.

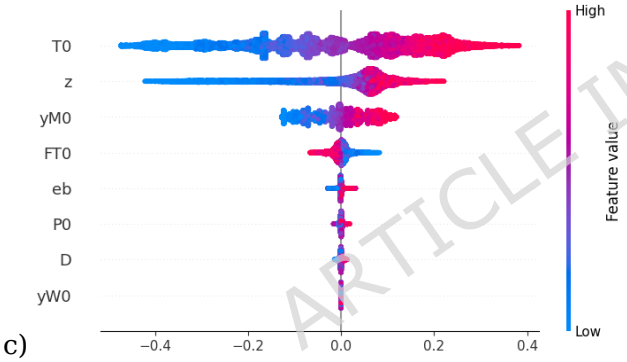
In plot (b), corresponding to reactor pressure (P), the most influential variable is P_0 (initial pressure). As expected from fundamental fluid dynamics, higher feed pressure (red) increases the outlet pressure, leading to high positive SHAP values. The second most important variable is z, though its influence is relatively complex: lower values of z (blue points) are generally associated with lower SHAP values, indicating pressure drop along the reactor. This matches the physical expectation that longer reactor lengths typically result in more pressure loss due to friction and reaction progression. The SHAP plots also reveal that some variables, such as D, eb, and y_{W0}, have minor and less consistent effects on the model outputs, evidenced by their narrow SHAP value distributions centered around zero in almost all subplots.

Overall, the SHAP-based sensitivity analysis not only quantifies the impact of input variables on the prediction but also provides physically interpretable insights consistent with the underlying chemical process. This validates the model's learning behavior and supports its use for further scenario analysis or optimization studies.

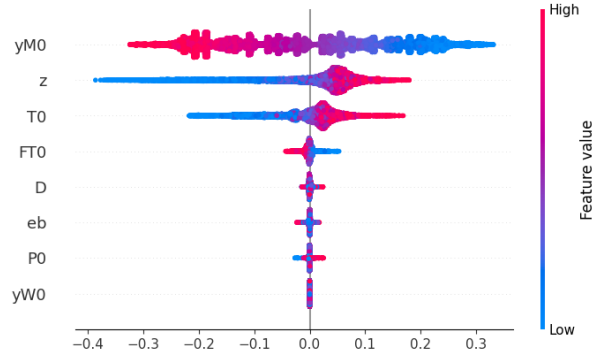


a) b)





c) d)



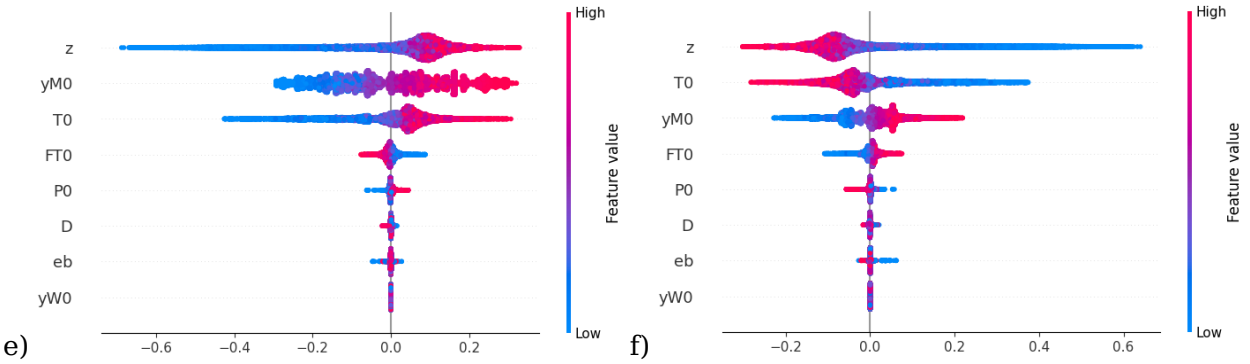


Fig.15. SHAP analysis of the input / output variables for data-driven model; a) Conversion, b) Pressure, c) Temperature, d) Mole fraction of water, e) Mole fraction of DME, f) Mole fraction of methanol

4. Conclusion

In this study, an interpretable hybrid modeling framework was successfully developed and demonstrated for the synthesis of dimethyl ether (DME) in a fixed-bed catalytic reactor. The framework effectively integrates first-principles modeling, data-driven learning, and optimization into a unified architecture capable of accurate prediction, physical interpretability, and computational efficiency.

A validated first-principles simulator was first established by solving the governing conservation equations for mass, energy, and momentum under realistic operating conditions. Based on this simulator, a comprehensive synthetic databank of 7000 samples was generated using the Latin Hypercube Sampling (LHS) technique, covering eight key input variables

catalyst bed length, inlet molar flow rate, initial temperature, initial pressure, methanol and water inlet concentrations, reactor diameter, and bed porosity and six output variables, including component concentrations, outlet temperature, outlet pressure, and conversion.

Among the data-driven models evaluated (XGB, KNN, and GBR), the XGB algorithm achieved superior accuracy with an average R^2 of 0.999 and minimal MSE across all outputs. Furthermore, two hybrid modeling approaches hybrid estimation (kinetic replacement) and hybrid correction (LSTM-based sequential extrapolation) were implemented to enhance predictive performance while maintaining physical interpretability. The LSTM-based correction model achieved an MSE of 4e-4, outperforming other methods and confirming its ability to accurately reproduce and improve upon first-principles predictions. These findings demonstrate that the proposed hybrid framework can both improve the predictive accuracy of first-principles models and replace complex kinetic sub-models with interpretable data-driven components.

In addition, by coupling the XGB model with the Differential Evolution (DE) optimization algorithm, optimal operating conditions were determined yielding a maximum conversion of 84.3% and a minimal temperature rise of 84.9 K. This integration highlights the framework's potential for process optimization, enabling efficient exploration of operational trade-offs between conversion and thermal behavior.

Overall, the developed interpretable hybrid modeling framework represents a generalizable and physically consistent approach that bridges mechanistic understanding and data-driven learning. It offers a scalable foundation for reactor design, control, and optimization in catalytic systems and other complex chemical processes.

Building upon the current findings, several future research directions are recommended:

- Substitute empirical models with data-driven representations by replacing the Peng–Robinson equation of state with machine learning-based thermodynamic estimators trained on experimental data;
- Explore advanced hybrid techniques, such as physics-informed neural networks, to enhance physical consistency and generalization under unseen conditions; and
- Recalibrate hybrid models using industrial or pilot-scale data to ensure reliable application under real operating conditions.

Through these extensions, the proposed methodology can evolve into a powerful tool for interpretable and optimization-driven modeling across a wide range of chemical engineering processes.

References

1. Lourenço, M.P., et al., An adaptive design approach for defects distribution modeling in materials from first-principle calculations. Journal of molecular modeling, 2020. **26**: p. 1-12.
2. Chun, H., et al., First-principle-data-integrated machine-learning approach for high-throughput searching of ternary electrocatalyst toward oxygen reduction reaction. Chem Catalysis, 2021. **1**(4): p. 855-869.
3. Zahedi, G., et al., Hybrid artificial neural network First principle model formulation for the unsteady state simulation and analysis of a packed bed reactor for CO2 hydrogenation to methanol. Chemical Engineering Journal, 2005. **115**(1-2): p. 113-120.
4. Schmidt, J., et al., Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 2019. **5**(1): p. 83.
5. Stein, A.F., et al., A hybrid modeling approach to resolve pollutant concentrations in an urban area. Atmospheric Environment, 2007. **41**(40): p. 9410-9426.
6. Kauwe, S.K., et al., Machine learning prediction of heat capacity for solid inorganics. Integrating Materials and Manufacturing Innovation, 2018. **7**: p. 43-51.
7. Bhutani, N., G. Rangaiah, and A. Ray, First-principles, data-based, and hybrid modeling and optimization of an industrial hydrocracking unit. Industrial & engineering chemistry research, 2006. **45**(23): p. 7807-7816.
8. Nazemzadeh, N., et al., Integration of first-principle models and machine learning in a modeling framework: An application to flocculation. Chemical Engineering Science, 2021. **245**: p. 116864.
9. Nielsen, R.F., et al., An uncertainty-aware hybrid modelling approach using probabilistic machine learning, in Computer Aided Chemical Engineering. 2021, Elsevier. p. 591-597.

10. Belyadi, H. and A. Haghighat, Machine learning guide for oil and gas using Python: A step-by-step breakdown with data, algorithms, codes, and applications. 2021: Gulf Professional Publishing.
11. Park, S., et al., Machine Learning Applications for Chemical Reactions. Chemistry–An Asian Journal, 2022. **17**(14): p. e202200203.
12. Dobbelaere, M.R., et al., Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering, 2021. **7**(9): p. 1201-1211.
13. Yan, Y., T. Borhani, and P. Clough, Machine learning applications in chemical engineering. 2020.
14. Carranza-Abaid, A. and J.P. Jakobsen, Neural network programming: Integrating first principles into machine learning models. Computers & Chemical Engineering, 2022. **163**: p. 107858.
15. Sharma, N. and Y. Liu, A hybrid science-guided machine learning approach for modeling chemical processes: A review. AIChE Journal, 2022. **68**(5): p. e17609.
16. Karniadakis, G.E., et al., Physics-informed machine learning. Nature Reviews Physics, 2021. **3**(6): p. 422-440.
17. Paszke, A., et al., Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 2019. **32**.
18. Abadi, M., et al. {TensorFlow}: a system for {Large-Scale} machine learning. in 12th USENIX symposium on operating systems design and implementation (OSDI 16). 2016.
19. McKinney, W. Data structures for statistical computing in Python. in SciPy. 2010.
20. Harris, C.R., et al., Array programming with NumPy. Nature, 2020. **585**(7825): p. 357-362.
21. Pedregosa, F., et al., Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011. **12**: p. 2825-2830.

22. Virtanen, P., et al., SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods, 2020. **17**(3): p. 261-272.
23. Sansana, J., et al., Recent trends on hybrid modeling for Industry 4.0. Computers & Chemical Engineering, 2021. **151**: p. 107365.
24. Zendehboudi, S., N. Rezaei, and A. Lohi, Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review. Applied energy, 2018. **228**: p. 2539-2566.
25. Bikmukhametov, T. and J. Jäschke, Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. Computers & Chemical Engineering, 2020. **138**: p. 106834.
26. Bradley, W., et al., Perspectives on the integration between first-principles and data-driven modeling. Computers & Chemical Engineering, 2022. **166**: p. 107898.
27. Sun, B., et al., A comprehensive hybrid first principles/machine learning modeling framework for complex industrial processes. Journal of Process Control, 2020. **86**: p. 30-43.
28. Nasiri, P. and R. Dargazany, Reduced-PINN: An Integration-Based Physics-Informed Neural Networks for Stiff ODEs. arXiv preprint arXiv:2208.12045, 2022.
29. Ji, W., et al., Stiff-pinn: Physics-informed neural network for stiff chemical kinetics. The Journal of Physical Chemistry A, 2021. **125**(36): p. 8098-8106.
30. Jinnouchi, R., F. Karsai, and G. Kresse, Making free-energy calculations routine: combining first principles with machine learning. Physical Review B, 2020. **101**(6): p. 060201.
31. Veit, M., et al., Equation of state of fluid methane from first principles with machine learning potentials. Journal of chemical theory and computation, 2019. **15**(4): p. 2574-2586.

32. Schäfer, P., et al., The Potential of Hybrid Mechanistic/Data-Driven Approaches for Reduced Dynamic Modeling: Application to Distillation Columns. Chemie Ingenieur Technik, 2020. **92**(12): p. 1910-1920.
33. Rodriguez, C., P. Mhaskar, and V. Mahalec, Linear hybrid models of distillation towers. Computers & Chemical Engineering, 2023. **171**: p. 108160.
34. Di Caprio, U., et al., Predicting overall mass transfer coefficients of CO2 capture into monoethanolamine in spray columns with hybrid machine learning. Journal of CO2 Utilization, 2023. **70**: p. 102452.
35. Dong, S., Y. Zhang, and X. Zhou, Intelligent Hybrid Modeling of Complex Leaching System Based on LSTM Neural Network. Systems, 2023. **11**(2): p. 78.
36. Khalid, R.Z., et al., Comparison of Standalone and Hybrid Machine Learning Models for Prediction of Critical Heat Flux in Vertical Tubes. Energies, 2023. **16**(7): p. 3182.
37. Yang, Q., et al., A hybrid data-driven machine learning framework for predicting the performance of coal and biomass gasification processes. Fuel, 2023. **346**: p. 128338.
38. dos Santos Junior, J.M., Í.A.M. Zelioli, and A.P. Mariano, Hybrid Modeling of Machine Learning and Phenomenological Model for Predicting the Biomass Gasification Process in Supercritical Water for Hydrogen Production. Eng, 2023. **4**(2): p. 1495-1515.
39. Ren, S., S. Wu, and Q. Weng, Physics-informed machine learning methods for biomass gasification modeling by considering monotonic relationships. Bioresource Technology, 2023. **369**: p. 128472.
40. Tsopanoglou, A. and I.J. del Val, Moving towards an era of hybrid modelling: advantages and challenges of coupling mechanistic and data-driven models for upstream pharmaceutical bioprocesses. Current Opinion in Chemical Engineering, 2021. **32**: p. 100691.
41. Cheng, Z., A. Ronen, and H. Yuan, Hybrid Modeling of Engineered Biological Systems through Coupling Data-Driven Calibration of Kinetic

Parameters with Mechanistic Prediction of System Performance. bioRxiv, 2023: p. 2023.06. 14.545039.

42. Zahedi, G., A. Lohi, and K. Mahdi, Hybrid modeling of ethylene to ethylene oxide heterogeneous reactor. Fuel processing technology, 2011. **92**(9): p. 1725-1732.
43. Luo, N., et al., Development of a hybrid model for industrial ethylene oxide reactor. Industrial & engineering chemistry research, 2012. **51**(19): p. 6926-6932.
44. Bui, L., et al., A Hybrid Modeling Approach for Catalyst Monitoring and Lifetime Prediction. ACS Engineering Au, 2021. **2**(1): p. 17-26.
45. Riyono, B., et al., A hybrid machine learning approach for improving fuel temperature prediction of research reactors under mix convection regime. Results in Engineering, 2022. **15**: p. 100612.
46. Kordkheili, M.S. and F. Rahimpour, Artificial neural network and semi-empirical modeling of industrial-scale Gasoil hydrodesulfurization reactor temperature profile. Mathematics and Computers in Simulation, 2023. **206**: p. 198-215.
47. Mehrani, M.-J., et al., Application of a hybrid mechanistic/machine learning model for prediction of nitrous oxide (N2O) production in a nitrifying sequencing batch reactor. Process Safety and Environmental Protection, 2022. **162**: p. 1015-1024.
48. Li, K., et al., An integrated first principal and deep learning approach for modeling nitrous oxide emissions from wastewater treatment plants. Environmental Science & Technology, 2022. **56**(4): p. 2816-2826.
49. Murakami, Y. and A. Shono, Reaction engineering with recurrent neural network: Kinetic study of Dushman reaction. Chemical Engineering Journal Advances, 2022. **9**: p. 100219.
50. Lan, T. and Q. An, Discovering catalytic reaction networks using deep reinforcement learning from first-principles. Journal of the American Chemical Society, 2021. **143**(40): p. 16804-16812.

51. Ghosh, D., et al., Hybrid modeling approach integrating first-principles models with subspace identification. Industrial & Engineering Chemistry Research, 2019. **58**(30): p. 13533-13543.
52. Hassanpour, H., P. Mhaskar, and M.J. Risbeck, A hybrid machine learning approach integrating recurrent neural networks with subspace identification for modelling HVAC systems. The Canadian Journal of Chemical Engineering, 2022. **100**(12): p. 3620-3634.
53. Patel, R., S. Bhartiya, and R. Gudi, Optimal temperature trajectory for tubular reactor using physics informed neural networks. Journal of Process Control, 2023. **128**: p. 103003.
54. Azarpour, A., et al., A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors. Chemical Engineering Research and Design, 2017. **117**: p. 149-167.
55. Azarpour, A., et al., Catalytic activity evaluation of industrial Pd/C catalyst via gray-box dynamic modeling and simulation of hydropurification reactor. Applied Catalysis A: General, 2015. **489**: p. 262-271.
56. Peterson, L., J. Bremer, and K. Sundmacher, Hybrid modeling of the catalytic CO2 methanation using process data and process knowledge, in Computer Aided Chemical Engineering. 2023, Elsevier. p. 1489-1494.
57. Delgado Otalvaro, N., et al., Kinetics of the Direct DME Synthesis: State of the Art and Comprehensive Comparison of Semi-Mechanistic, Data-Based and Hybrid Modeling Approaches. Catalysts, 2022. **12**(3): p. 347.
58. Ammar, Y., P. Cognet, and M. Cabassud, ANN for hybrid modelling of batch and fed-batch chemical reactors. Chemical Engineering Science, 2021. **237**: p. 116522.
59. Bakhtyari, A., M. Mofarahi, and A. Iulianelli, Combined mathematical and artificial intelligence modeling of catalytic bio-methanol conversion to dimethyl ether. Energy Conversion and Management, 2023. **276**: p. 116562.

60. Ng, A., Machine learning. coursera. Standford University, 2016.
61. Dangeti, P., Statistics for machine learning. 2017: Packt Publishing Ltd.
62. Géron, A., Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. 2022: " O'Reilly Media, Inc.".
63. James, G., et al., An introduction to statistical learning. Vol. 112. 2013: Springer.
64. Chen, T. and C. Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
65. Peterson, L.E., K-nearest neighbor. Scholarpedia, 2009. **4**(2): p. 1883.
66. Sherstinsky, A., Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 2020. **404**: p. 132306.
67. Hochreiter, S. and J. Schmidhuber, Long Short-Term Memory. Neural Computation, 1997. **9**(8): p. 1735-1780.
68. Kamath, C., Intelligent sampling for surrogate modeling, hyperparameter optimization, and data analysis. Machine Learning with Applications, 2022. **9**: p. 100373.
69. Prieto, A., M. Atencia, and F. Sandoval, Advances in artificial neural networks and machine learning. 2013, Elsevier. p. 1-4.
70. Steponavičė, I., et al., On sampling methods for costly multi-objective black-box optimization. Advances in stochastic and deterministic global optimization, 2016: p. 273-296.
71. Bashiri, S., E. Yasari, and S. Tayyebi, Comparison of different sampling and surrogate modelling approaches for a multi-objective optimization problem of direct dimethyl ether synthesis in the fixed-bed reactor. Chemometrics and Intelligent Laboratory Systems, 2022. **230**: p. 104683.
72. Ansari, S., et al., Prediction of hydrogen solubility in aqueous solutions: Comparison of equations of state and advanced machine learning-

metaheuristic approaches. International Journal of Hydrogen Energy, 2022. **47**(89): p. 37724-37741.

73. Cohen, I., et al., Pearson correlation coefficient. Noise reduction in speech processing, 2009: p. 1-4

74. L. M. Rios and N. V. Sahinidis, "Derivative-free optimization: a review of algorithms and comparison of software implementations," *Journal of Global Optimization,* vol. 56, pp. 1247-1293, 2013.

75. R. Storn and K. Price, "Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization,* vol. 11, no. 4, p. 341, 1997.

76. D. Kraft, "A software package for sequential quadratic programming," *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt,* 1988.

77. R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM Journal on Optimization,* vol. 9, no. 4, pp. 877-900, 1999.

78. S. M. Lundberg, S. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

Funding

No funding was received for this study.

Data availability

All data generated or analysed during this study are included in supplementary information files.