Scilight

Smart Chemical Engineering

https://www.sciltp.com/journals/sce

Article

Data-Driven Modeling and Bayesian Optimization for the Formaldehyde-Acetylene Reaction in a Slurry Bed Reactor

Xiao-Qi Liu [1], Zuo-Qian Jihou [1], Hui-Long Wei [1,*] and Zheng-Hong Luo [1,2,*]

[1] State Key Laboratory of High-Efficiency Utilization of Coal and Green Chemical Engineering, College of Chemistry and Chemical Engineering, Ningxia University, Yinchuan 750014, China

[2] Department of Chemical Engineering, School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China

* Correspondence: weihl@nxu.edu.cn (H.-L.W.); luozh@sjtu.edu.cn (Z.-H.L.)

How To Cite: Liu, X.-Q.; Jihou, Z.-Q.; Wei, H.-L.; et al. Data-Driven Modeling and Bayesian Optimization for the Formaldehyde-Acetylene Reaction in a Slurry Bed Reactor. *Smart Chemical Engineering* **2025**, *1*(1), 7. https://doi.org/10.53941/sce.2025.100007

Received: 29 August 2025
Revised: 13 November 2025
Accepted: 1 December 2025
Published: 5 December 2025

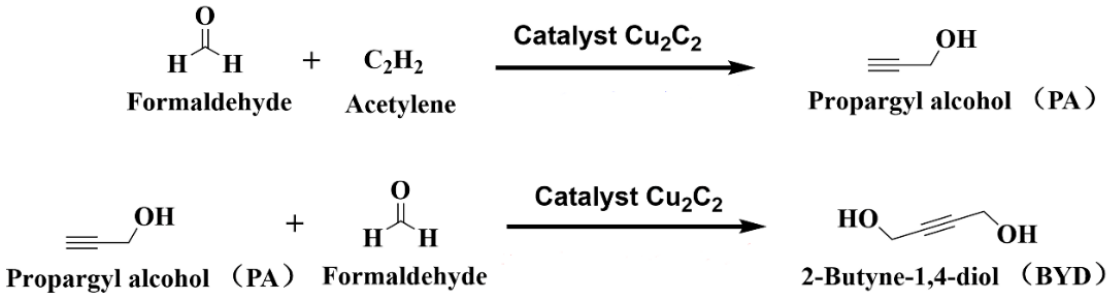
Abstract: 1,4-Butynediol (BYD), an essential intermediate for fine chemicals and polymer production, is primarily synthesized via formaldehyde-acetylene reaction. Kinetic experiments were conducted in a quasi-industrial slurry bed reactor under the conditions of 55–85 °C, 0–10.5 h, and pH 5–9 to obtain the time-resolved yield of 1,4-butynediol and the conversion of formaldehyde. The experimental results revealed that pH was a critical influencing factor on reaction performance, while it can not be directly coupled into mechanistic kinetic models. Therefore, four machine learning models, i.e., random forest (RF), extremely randomized trees (Extra Trees, ET), light gradient boosting machine (LightGBM) and extreme gradient boosting (XGBoost) were employed to establish data-driven models that can directly capture the pH influence. The 84 experimental data points were augmented to 1023 samples by interpolation and extrapolation method, then the dataset was split into training, validation, and testing subsets in a 6:2:2 ratio. The training results demonstrated that the XGBoost model exhibited the best generalization ability and stability, achieving the highest average coefficient of determination (R^2) for formaldehyde conversion (0.9847 ± 0.0022) and 1,4-butynediol yield (0.9773 ± 0.0035), and the mean absolute error for both targets was less than 0.027. Finally, the XGBoost model was coupled with Bayesian optimization to search the optimal process parameters.

Keywords: 1,4-butynediol; machine learning; small-sample data; kinetic modeling; multi-objective optimization

1. Introduction

1,4-Butynediol (BYD) is an important intermediate that has attracted considerable attention due to its potential to be transformed into various high value-added chemicals. Through dehydration and dehydrogenation reactions, BYD can be converted into poly butylene terephthalate (PBT), γ-butyrolactone (GBL), polyurethane (PU), N-methylpyrrolidone (NMP), and poly butylene succinate (PBS), which are widely applied in the production of polymer materials, bioplastics, and biodegradable plastics [1,2]. In addition, BYD serves as a crucial synthetic precursor for vitamin A, vitamin B6, as well as a variety of pesticides and corrosion inhibitors [3–5]. The industrial production mainly relies on Reppe method, in which the copper-based catalysts promote the nucleophilic addition of acetylene to formaldehyde (HCHO), followed by further transformation of the propargyl alcohol intermediate into BYD (as shown in Scheme 1) [6–15]. The yield of BYD is strongly influenced by temperature and pH. Under alkaline conditions, formaldehyde undergoes a disproportionation reaction to produce formic acid and methanol, and the formic acid may further esterification with propargyl alcohol or BYD, results in the reduction in BYD

yield [16,17]. The industrial plant for BYD production is slurry bed reactor (SBR), in which the formaldehyde aqueous solution is mixed with the solid Cu-based catalyst, and acetylene is introduced into the liquid phase to initiate the reaction [7,18,19]. For this complex muti-phase reaction, systematic modeling and optimization can help improve the yield of BYD.



Scheme 1. The pathway of formaldehyde-acetylene reaction.

The classical kinetic modeling of complex reactions relies on mechanistic equations and the fitting of rate parameters. However, under conditions of sparse experimental data, such approaches are time-consuming, labor-intensive and often fail to accurately capture the real reaction process [20–22]. Previous studies have shown that although pH has critical effect on reaction rate, it is difficult to be directly incorporated into classical kinetic equations, thereby limiting the applicability of such models [23,24]. The advancement of machine learning (ML) algorithms has provided a novel pathway for modeling complex reaction processes, and numerous references have been reported in exploring this direction [25]. Edward et al. applied machine learning clustering to characterize the complex reaction mechanism of aldehyde combustion, identifying chemically reasonable reaction stages via local sensitivity analysis [26]. Matthew and Pierre developed 'ReactionPredictor', a machine-learning tool for mechanistic-level prediction of complex chemical reactions with high precision across reaction types and pathways [27].

Therefore, we introduced data-driven machine learning approaches to address the limitations of mechanistic models. The ML models can capture nonlinear features and reduce the dependence on complete mechanistic knowledge, thereby serving as an important complement to mechanism-based modeling [28–31]. Many researchers have adopted multi-model comparison strategies to evaluate predictive performance and applicability under practical conditions [31–33]. Such comparative studies highlighted differences among algorithms in terms of generalization, stability, and interpretability, and provided valuable guidance for model selection. Inspired by this, the present study conducted a comparative evaluation of four models within the data-augmented reaction system, i.e., RF, ET, LightGBM and XGBoost. Among the four models, XGBoost is endowed with inherent flexibility, robust nonlinear fitting capabilities, reliable generalization performance, and efficient training speed, making it widely adopted across disciplines such as chemistry, pharmaceutics, and materials science. In numerous prior studies, it has also been repeatedly validated as a high-performing machine learning approach, particularly for addressing complex predictive tasks in these domains [34,35].

In practical applications, experimental data remain limited due to difficulties in sensor deployment, high experimental costs, and noise interference, which can easily lead to overfitting in machine learning models and restrict their generalization ability [36–38]. This is particularly true for complex multi-phase reactions, such as those occurring in SBR, where the need for high catalyst dispersion and stable phase interactions further complicates the data collection process. The high cost of maintaining controlled reactor conditions, coupled with the time-consuming nature of experimentation, results in small sample sizes that hinder accurate model training [39–41]. Against this background, data augmentation techniques have been proposed to overcome the bottleneck of small-sample modeling [41,42]. By integrating physical mechanisms with statistical laws, data augmentation can generate virtual samples consistent with industrial logic, effectively compensating for insufficient parameter space coverage in experimental datasets [42–45], while reducing the risk of overfitting and thus enhancing the predictive accuracy and stability of models [46–48]. In this work, the augmented data generated under mixed experimental constraints are employed as the foundation for data-driven modeling.

Bayesian optimization can rapidly screen optimal solutions among multiple reaction pathways and conditions, significantly improving optimization efficiency and showing remarkable advantages across various fields [39–49]. Meanwhile, machine learning integrated with Bayesian optimization further promoted the efficient optimization of reaction conditions. Zhang et al. applied Bayesian optimization to maximize photoluminescence intensity in the growth process of tungsten disulfide (WS_2), achieving an 86.6% improvement after 13 iterations [50]. Wang et al. combined CatBoost with Bayesian optimization to optimize the catalytic ozonation process of antibiotics,

attaining high predictive accuracy (R^2 = 0.9482) and providing a novel approach for process design [51]. Building on this foundation, the present work systematically compares multiple machine learning models and data partitioning strategies, identifies the best-performing model (XGBoost) and optimal partitioning method, and further integrates them with Bayesian optimization to achieve effective multi-objective optimization of reaction conditions and comprehensive improvement in prediction accuracy.

This study focuses on the synthesis of BYD by formaldehyde-acetylene reaction, with particular attention to the influence of pH on reaction selectivity and yield. Under the constraint of limited experimental data, we proposed a physics-informed data augmentation strategy, which is further integrated with an XGBoost regression model and Bayesian optimization for modeling and optimization. On the data level, a strictly physically consistent interpolation–extrapolation network was constructed. After data augmentation, two sets of experimental conditions were deliberately excluded from model training and reserved for assessing the model's extrapolation capability. On the algorithmic level, four machine learning models, including RF, ET, LightGBM and XGBoost, were systematically evaluated under five different random seeds for parallel comparison. The model with the best generalization ability and result stability was selected as the prior model. After training, it was combined with Bayesian optimization–Gaussian process regression (BO-GPR) to realize multi-objective optimization for the yield of BYD and conversion of HCHO.

2. Experimental Section

2.1. Materials

Propargyl alcohol (99%, AR) was purchased from Chengdu McCarthy Chemical Co., Ltd. (Chengdu, China). 1,4-Butynediol (98%, AR) was obtained from Shanghai Aladdin Biochemical Technology Co., Ltd. (Shanghai, China). Formaldehyde (37%, AR) was supplied by Macklin Biochemical Co., Ltd. (Shanghai, China). Sodium sulfite anhydrous (99.5%, AR), sodium hydroxide (96%, AR), and ethanol (99.7%, AR) were purchased from Sinopharm Chemical Reagent Co., Ltd. (Shanghai, China). Thymolphthalein (99%, AR) was purchased from Shanghai Yien Chemical Technology Co., Ltd. (Shanghai, China). Standardized sulfuric acid titration solution (0.1 mol/L, 99%, AR) was supplied by Guangzhou Howe Pharmaceutical Technology Co., Ltd. (Guangzhou, China). Purified water was obtained from Hangzhou Wahaha Group Co., Ltd. (Hangzhou, China). Acetylene and high-purity nitrogen were purchased from Ningxia Guangli Gas Co., Ltd. (Ningxia, China). Basic copper bismuth carbonate catalyst was purchased from China Petrochemical Great Wall Energy and Chemical Co., Ltd. (Ningxia, China). All the chemical reagents were used directly in the experiments without further purification.

2.2. Experimental Setup

The Programmable Logic Controller (PLC) controlled SBR system is shown in Figure 1. This system integrates PLC and computer operation, enabling the control and display of temperature, flow, and pressure. The core equipment for the kinetic experiments is a jacket-heated 316L high-pressure reactor (volume 2 L, inner diameter 120 mm, height 400 mm) manufactured by Xi'an Shiyerui Scientific Equipment Co., Ltd. (Xi'an, China), in which a 200-mesh stainless-steel wire mesh is installed to filter the catalyst.

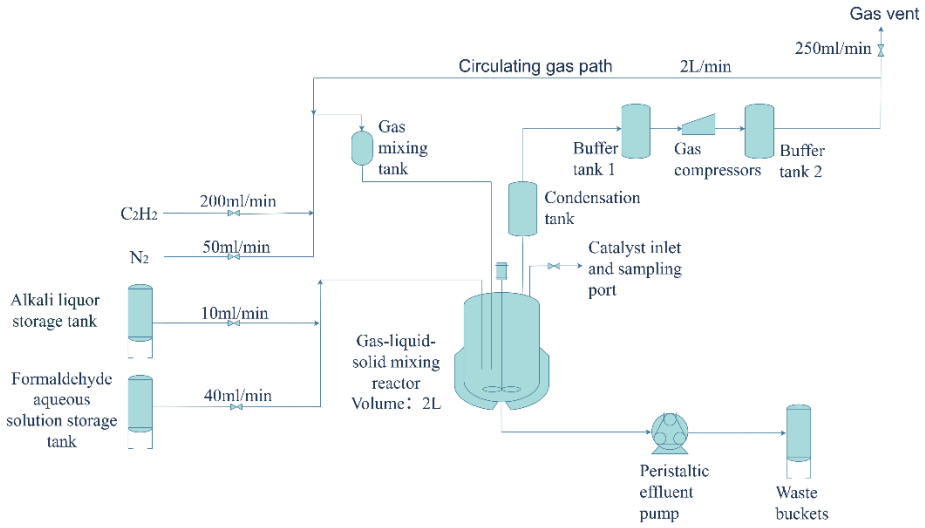


Figure 1. The simplified flow chart of the PLC SBR system customized in this work.

After the catalyst is introduced into the reactor, nitrogen protective gas is fed from the gas cylinder via a pressure-reducing valve, then passes through a two-way ball valve and a gas filter, before being directed into a mass flow controller for flow regulation. Subsequently, the gas passes through a check valve into a gas-mixing tank, and is finally introduced into SBR. After the air inside the reactor is completely purged, acetylene gas is introduced following the same procedure. The heating temperature of the external jacket is set to control the reaction temperature. A micro-metering pump is used to deliver a quantitative amount of aqueous formaldehyde solution into a stainless-steel infusion tube, which then passes through a pressure gauge and a check valve before being injected into SBR. The alkaline solution is delivered into the infusion line by another micro-metering pump, which is interlocked with a pH meter to regulate the pH value. The solution also passes through a pressure gauge and check valve before entering the reactor. An anchor-type agitator is then started, and its speed is adjusted to initiate the ethynylation reaction. The actual temperature inside the reactor is monitored by sensors, which in turn adjust the jacket heating system to ensure precise temperature control with an accuracy of ±0.2 °C. The internal pressure is regulated and stabilized by controlling the pressure-reducing valve and the gas flow meter. In addition, the system is equipped with a gas condensation–circulation and venting unit. The overall system pressure is stabilized by the combined operation of a coil condenser, a gas compressor, and a gas buffer tank.

During the reaction, samples are collected at specific intervals through the catalyst inlet using a disposable syringe connected with Polytetrafluoroethylene tubing for subsequent analysis. After the reaction is completed, the spent liquid is pumped out using a discharge pump and transferred into a waste storage tank. The kinetic experiments are run under the following conditions: temperature range of 55–85 °C, pH range of 5–9 and reaction time of 0–10.5 h.

2.3. Analysis Method

The concentration of HCHO in the reaction liquid was determined using the titration method with anhydrous sodium sulfite, detailed operating procedures can be found in the Supplementary Materials.

The concentrations of BYD and propargyl alcohol (PA) were determined by gas chromatography (GC) with 1, 4-butanediol as the internal standard component, and the standard curves of PA and BYD are put in Figures S1 and S2 in the Supplementary Materials. GC analysis was performed on an Agilent 7820 system equipped with a DB-WAX (Agilent, Santa Clara, CA, USA) column (30 m × 0.32 mm × 0.25 μm, Agilent Technologies (China) Co., Ltd. Beijing, China), which can accurately analyze the content of each component in the reaction products. A flame ionization detector (FID) was employed owing to its low detection limit and high sensitivity toward organic compounds. Prior to analysis, the instrument was thoroughly calibrated to ensure measurement accuracy and operational reliability. The operating conditions were as follows: carrier gas, N_2; initial column temperature, 50 °C (held for 2 min); ramped from 50 °C to 183 °C at 30 °C min^{-1} (held for 0.5 min); then from 183 °C to 200 °C at 30 °C min^{-1} (held for 0.5 min); and finally from 200 °C to 240 °C at 30 °C min^{-1} (held for 1 min). The detector temperature was set at 280 °C, and the injector temperature at 270 °C. The reaction liquid sample was diluted with 50 mL of ethanol and then entered for GC analysis and the single injection volume was set to 0.2 μL, with an auto-sampler employed for injection. Each sample was injected three times, and the final result was determined using the average value of these three injections. The conversion of HCHO was calculated by:

$$X = \frac{C^0_{\mathrm{HCHO}} - C^t_{\mathrm{HCHO}}}{C^0_{\mathrm{HCHO}}} \tag{1}$$

The yield of BYD can be calculated by:

$$Y = \frac{2C^t_{\mathrm{BYD}}}{C^0_{\mathrm{HCHO}}} \tag{2}$$

where C^0_{HCHO}, C^t_{HCHO} and C^t_{BYD} are the initial concentration of HCHO, the concentration of HCHO at sampling time *t*, and the concentration of BYD at sampling time *t*, respectively.

3. Modeling Approach

3.1. ML Model

XGBoost is an ensemble algorithm that uses regression trees as base learners. Its fundamental approach involves employing gradient boosting to progressively optimize each weak learner. During each iteration, it fits and updates the residuals from the previous round's predictions, combining multiple weak learners into a single strong learner. Compared to traditional boosting tree methods, XGBoost introduces L2 regularization into its

objective function to prevent overfitting and further enhances model efficiency and accuracy by expanding the loss function using a second-order Taylor approximation. The principle of the XGBoost algorithm is illustrated in Figure 2. Its objective function is expressed as:

$$Y(\varphi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) = \sum_{i=1}^{k} \Omega(f_k,) \tag{3}$$

where $Y(\varphi)$ is the objective function, l is the loss function, y_i and $\hat{y}_i$ are the measured value and predicted value of sample i, respectively; n is the number of samples, Ω is the regularization term; f_k is the complexity of the k-th tree; k is the number of decision trees.

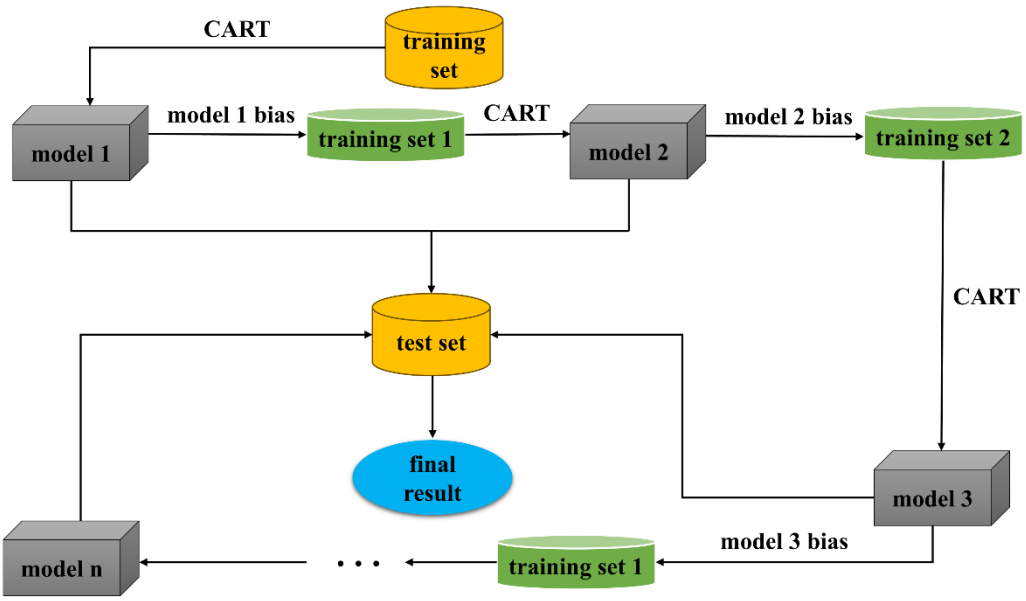


Figure 2. The principle of the XGBoost algorithm. (CART: classification and regression tree)

In terms of model estimation, the coefficient of determination (R^2) and mean absolute error (MAE) are used to quantify the prediction accuracy of the model. R^2 is used to measure the degree of fit of the model to the data, as shown in Equation (4). The closer R^2 is to 1, the higher the degree of agreement between the prediction results and the experimental results. MAE is used to measure the difference between the predicted values and the actual observed values, as shown in Equation (5). Complementing each other, the two can more comprehensively reflect the prediction performance of the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{\text{exp}} - y_{\text{predict}})^2}{\sum_{i=1}^{n}(y_{\text{exp}} - \bar{y}_{\text{exp}})^2} \tag{4}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_{\text{exp}} - \bar{y}_{\text{predict}}| \tag{5}$$

where y_{predict} is the predicted value of the model, $\bar{y}_{\text{exp}}$ and $\bar{y}_{\text{predict}}$ are the average values of the experimental values and predicted values respectively.

3.2. Model Framework

The temperature, reaction time, and pH were defined as input parameters, the conversion of HCHO and the yield of BYD were defined as output variables. Based on the experimental dataset, we introduced strict physicochemical constraints and applied grid-based augmentation to expand the dataset. During the model construction stage, four commonly used machine learning algorithms, including RF, ET, LightGBM and XGBoost, were evaluated. Each model was trained and tested under five different random seeds, and the best-performing one was selected as the prior model. This model was then integrated with Bayesian optimization based on Gaussian process regression (BO-GPR). In the optimization process, 15 initial sampling points were adopted, followed by 80 iterations under a fixed random seed of 42, ensuring both stability and reproducibility of the search results.

Through this bi-objective optimization framework, we systematically explored the effects of reaction conditions on HCHO conversion and BYD yield, and ultimately identified the optimal operating conditions. The workflow: experimental data augmentation, multi-model comparative modeling and Bayesian optimization establishes a complete closed loop from data preparation and model construction to reaction conditions optimization.

4. Results and Discussion

4.1. Experimental Results

At pH values of 5, 7 and 9, the experimental variations of HCHO conversion and BYD yield with reaction time under different reaction temperatures are shown in Figure 3. Before the experiment began, three parallel trials were conducted under a selected condition to assess the repeatability of the system. The errors for each trial were found to be within a range of 5%, confirming the stability and reliability of the data obtained from the equipment. These results have been validated and are included in the Supplementary Materials, Figure S3.

In addition to the general trend that the reactant conversion increases with prolonged reaction time and the product yield gradually rises, the experimental data also exhibit several noteworthy phenomena. Under identical temperature and reaction time conditions, the yield of BYD under neutral conditions (Figure 3E) is significantly higher than that under either acidic or alkaline conditions (Figure 3D,F). Moreover, due to the presence of the Cannizzaro reaction, the decline of formaldehyde concentration proceeds more rapidly under alkaline conditions (Figure 3F). These observations indicate that the pH value plays a crucial regulatory role in reaction results, and thus must be considered as a key influencing factor in kinetic modeling. In addition, reaction time and temperature, as fundamental determinants of reaction kinetics and thermodynamics, are also incorporated into the data-driven model to ensure comprehensiveness and reliability of the results.

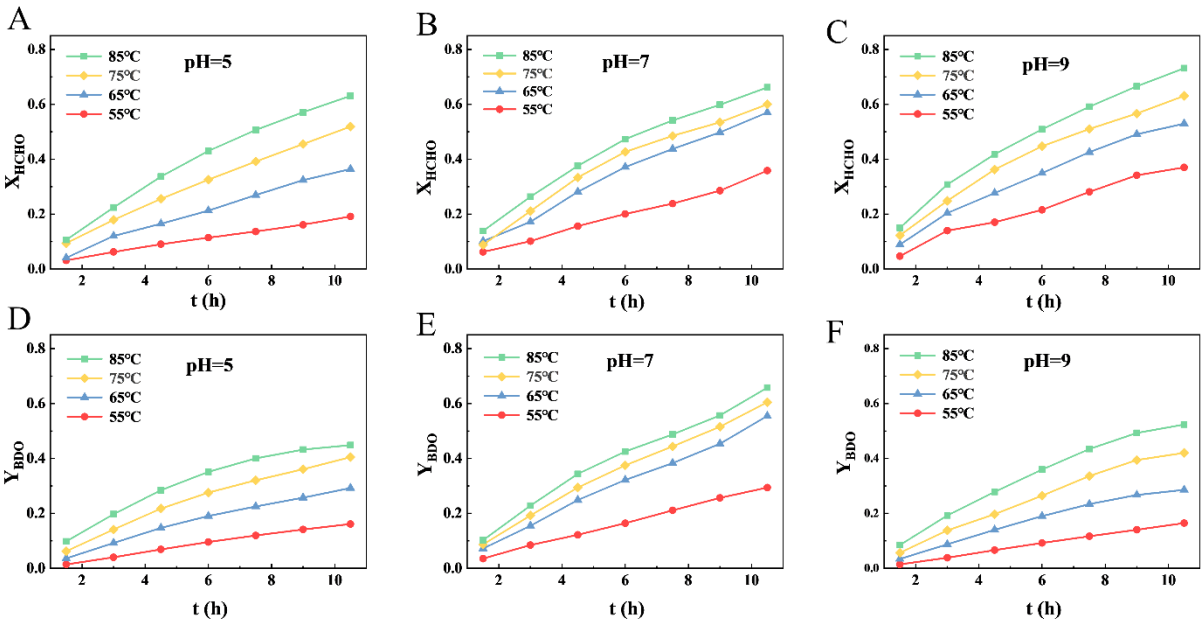


Figure 3. (**A**–**C**) Time-dependent variation of HCHO conversion at different temperatures under pH = 5, 7 and 9, (**D**–**F**) Time-dependent variation of BYD yield at different temperatures under pH = 5, 7 and 9.

4.2. Data Augmentation

To broaden the reaction window (55–85 °C, pH 5–9, 0–15.5 h) and enhance the model's predictive reliability, data augmentation was carried out based on the experimental dataset. The adopted strategies and their corresponding effects are summarized in Table 1. After augmentation, the original 84 experimental points were expanded to 1023 samples (excluding those reserved for extrapolation validation), thereby covering the complete time–temperature–pH grid. Figure 4 shows that the augmented data effectively filled the sparse regions in the temperature–pH space, resulting in a more uniform and continuous overall distribution. Furthermore, Figure 5A,B illustrate that the augmented dataset smoothly reproduces the kinetic trends of BYD yield and HCHO conversion, accurately reflecting the synergistic influence of temperature and pH on reaction performance. Although data augmentation can potentially introduce overfitting due to the artificial expansion of data volume, this risk was minimized through Gaussian noise filtering and the enforcement of monotonic physical constraints, which ensured

physicochemical plausibility. The consistency between the augmented and experimental data was visually confirmed through distribution comparison, which exhibited similar trends across temperature, time, and pH dimensions. Overall, these results confirm that the augmented dataset maintains both physical realism and statistical reliability, providing a solid foundation for subsequent machine-learning modeling and optimization.

Table 1. The summary of data augmentation strategies and effects across time, pH, temperature and global dimensions.

Dimension	Strategy	Effect
time	interpolation + smoothing	monotonic kinetics, finer resolution
pH	ML extrapolation + trend constraints	peak near neutral
temperature	ML extrapolation + prior knowledge	thermal promotion captured
all vars	gaussian noise + filtering	robustness, avoids rigidity
dataset	Iterative selection	conservation ensured, consistent distribution

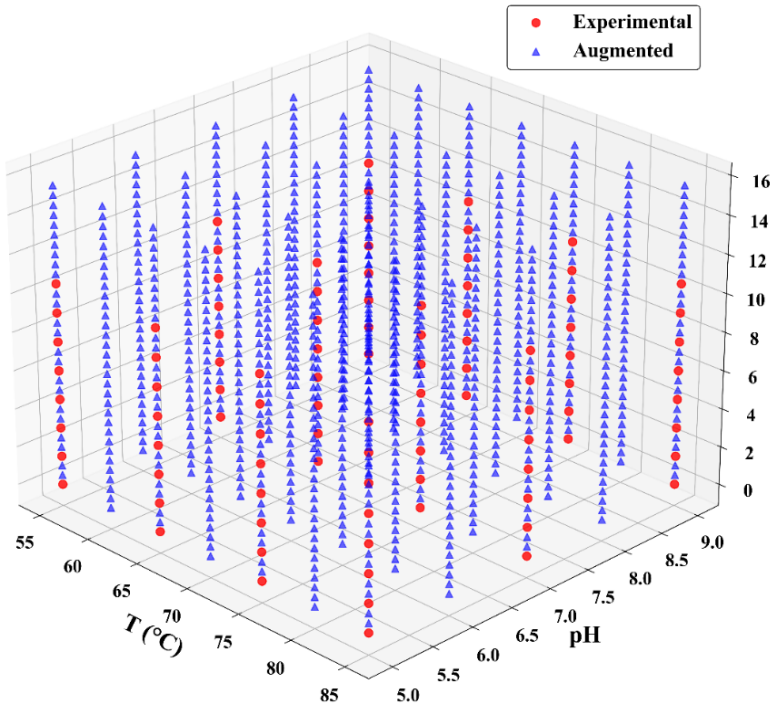


Figure 4. The 3D distribution map of experimental data and augmented data points.

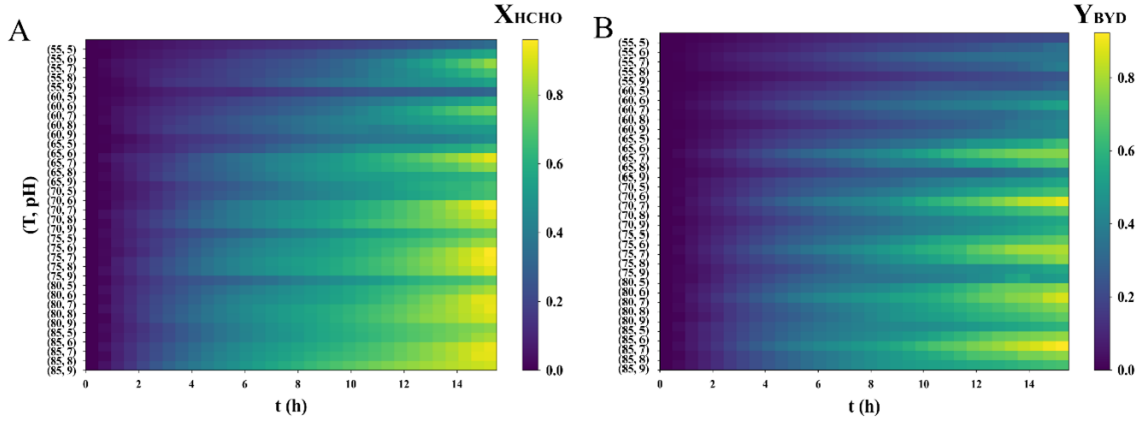


Figure 5. The heat map of (**A**) HCHO conversion and (**B**) BYD yield with enhanced data.

4.3. Model Estimation

We employed a ML modeling framework, beginning with a carefully designed parameterization tailored to the characteristics of chemical process data. A mixed splitting strategy was adopted for dataset partitioning, which incorporated both the original experimental data and the augmented dataset. The data were divided into training,

validation, and test sets in a ratio of 6:2:2. To avoid bias introduced by augmented data, stratified sampling was conducted based on two key process parameters, pH and temperature, thereby ensuring consistent proportions of experimental and augmented data across subsets and maintaining balanced and representative overall distributions.

To ensure a fair and systematic evaluation, four ensemble learning algorithms (RF, ET, XGBoost, and LightGBM) were implemented for model comparison. Their main hyperparameter settings are provided in Table 2. During model training, feature sampling, regularization, and physical constraints were jointly introduced to enhance generalization performance and ensure consistency with physical laws. Specifically, tree-based models applied 70–80% feature subsampling during node splitting, XGBoost further incorporated L1/L2 regularization to mitigate overfitting, while RF and ET achieved this through controlled tree depth and leaf size. Additionally, to ensure that the predictions adhered to reaction kinetics, non-negativity and monotonicity constraints were imposed so that HCHO conversion and BYD yield increased monotonically with time.

Table 2. The parameter settings of comparison models: main hyperparameters of RF, LightGBM, XGBoost, and ExtraTrees.

Model	Trees/Iterations	Max Deepth/Num Leaves	Learning Rate
XGBoost	200	6	0.05
RF	100	8	N/A
ET	100	8	N/A
LightGBM	200	31	0.05

As shown in Figure 6, the performance fluctuations of different models across runs were minimal, indicating strong overall stability of the modeling framework. Among the models, XGBoost demonstrated the best performance, with an average $R^2 = 0.9847 \pm 0.0022$ and MAE $= 0.0204 \pm 0.0010$ for predicting HCHO conversion, as well as $R^2 = 0.9773 \pm 0.0035$ and MAE $= 0.0207 \pm 0.0012$ for BYD yield. RF ranked second, with an average $R^2 = 0.9843 \pm 0.0036$ for HCHO conversion and $R^2 = 0.9693 \pm 0.0047$ for BYD yield. Although ExtraTrees and LightGBM showed slightly lower fitting accuracy, both still maintained reliable levels with $R^2 > 0.94$ and MAE < 0.04. Taken together, these findings demonstrate that XGBoost consistently outperformed the other models across both target variables. This superior performance can be attributed to its iterative gradient boosting framework and built-in regularization, which allow the model to capture complex nonlinear relationships between variables while effectively preventing overfitting. Additionally, through hypersensitive parameter analysis, the model can further fine-tune its settings for optimal performance. Unlike RF and ET, which build trees independently, XGBoost constructs trees sequentially based on residual errors, thereby improving predictive accuracy and ensuring greater stability. Consequently, XGBoost was selected as the core model for subsequent optimization analyses, with the random seed fixed at 42 to maintain reproducibility.

Furthermore, the generalization capability of the model was examined. Figure 7 presents the comparison between experimental values and XGBoost predictions. For both validation and test sets, the scatter points are closely distributed along the 45° diagonal line, indicating that the model performs well in both fitting accuracy and generalization ability.

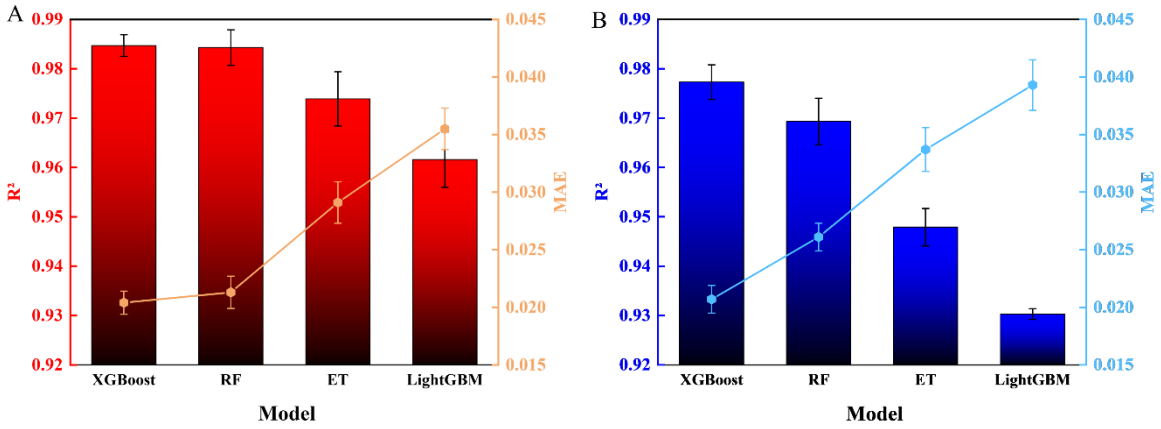


Figure 6. The comparison chart of ML model performance. (**A**) HCHO conversion; (**B**) BYD yield. (Solid bars denote R², whereas striped bars denote MAE.).

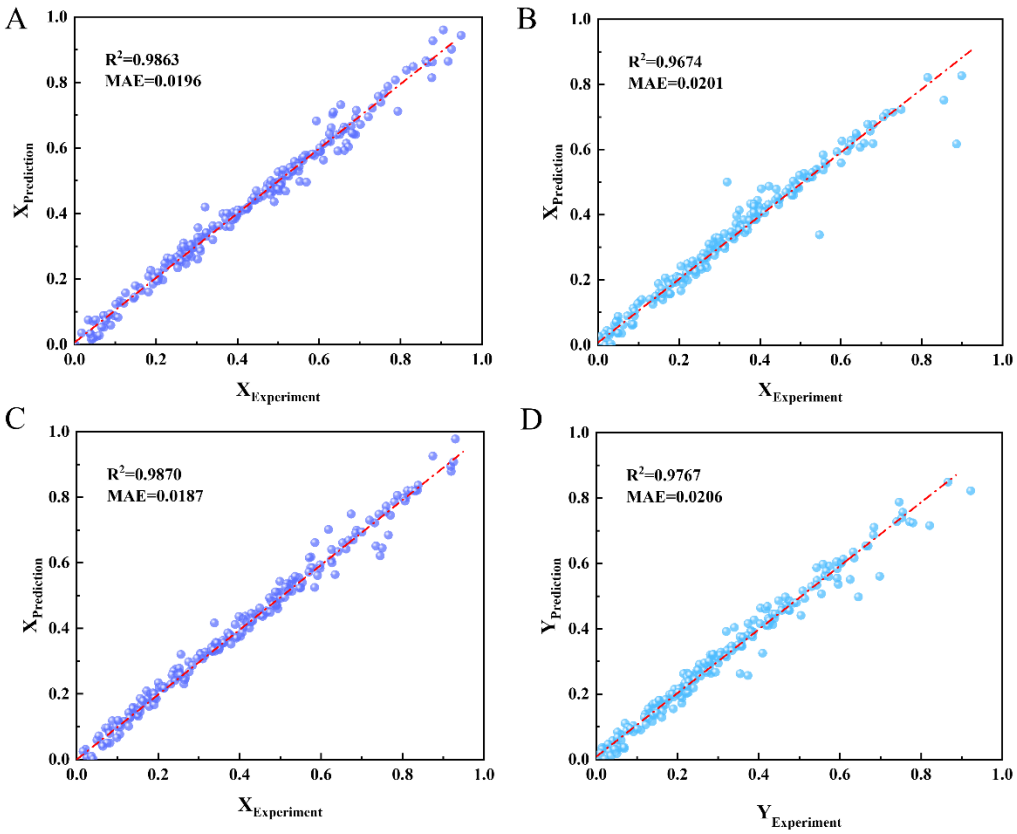


Figure 7. The predictive performance of (**A**) HCHO conversion in the test set; (**B**) BYD yield in the test set; (**C**) HCHO conversion in the validation set; (**D**) BYD yield in the validation set.

For a model, beyond its ability to predict data within the training dataset, the capability to accurately predict unseen data is an essential indicator of its scalability. To evaluate the model's applicability, we selected two sets of experimental data that were not involved in the modeling process. By examining the model's fitting performance on these datasets, we tested its generalization ability. As shown in Figure 8A,B, the model provides accurate predictions for these experimental points, demonstrating strong predictive performance in both conversion and yield.

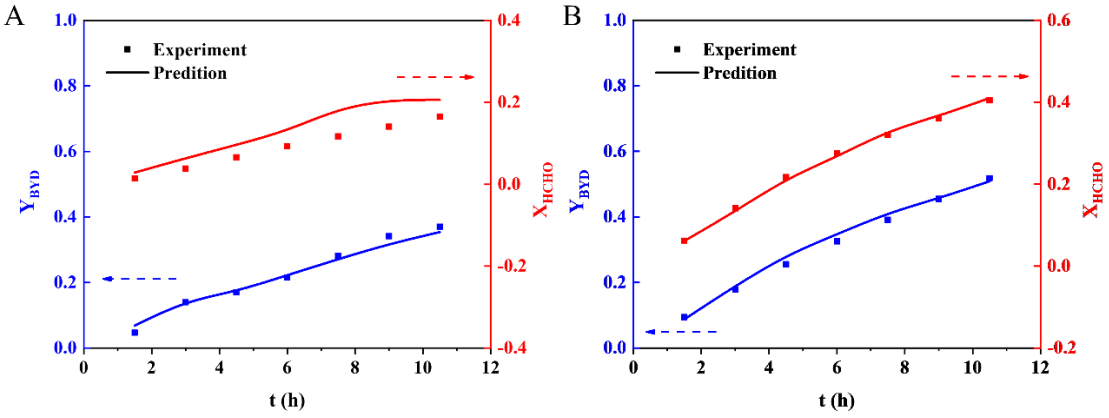


Figure 8. (**A**) Prediction performance of data not used for model training (pH = 5, T = 75 °C); (**B**) Prediction performance of data not used for model training (pH = 9, T = 55 °C).

In conclusion, XGBoost exhibited the best overall performance in terms of prediction accuracy, robustness, and generalization ability. This result further supports the rationality of our comparative analysis and the reliability of our model selection, providing a solid foundation for the application of this model in subsequent process design and optimization.

4.4. Bayesian Optimization

The construction of the kinetic model is intended to serve design calculations, which in industrial practice usually means further optimizing reaction conditions to achieve the highest possible conversion while meeting yield requirements. In the Reppe process for BYD synthesis, feed parameters such as formaldehyde concentration and acetylene partial pressure are typically fixed and cannot be easily adjusted. Therefore, this study focuses on analyzing the effects of pH, temperature, and time on HCHO conversion and BYD yield. To achieve global optimization of target yield and conversion within the multidimensional reaction condition space, Bayesian optimization is introduced to search and predict optimal experimental conditions. Specifically, the trained XGBoost model is employed as a surrogate model, combined with a gaussian process (GP) kernel function to capture predictive uncertainty. The optimization objective is defined as a multi-objective weighted function, where the BYD yield and HCHO conversion are assigned weights $w_1 = 0.7$ and $w_2 = 0.3$ respectively, thereby constructing the objective function:

$$f(x) = w_1 \cdot Y_{\mathrm{BYD}} + w_2 \cdot X_{\mathrm{HCHO}} \tag{6}$$

These weights were selected based on industrial operation priorities emphasizing higher BYD yield for process efficiency. Such a weighted objective function enables simultaneous consideration of yield and conversion, rather than optimizing them in isolation. A balance between yield and conversion is ensured to reflect industrial feasibility. The optimization is carried out within the three-dimensional search space of pH–temperature–time. During the optimization process, 15 initial random sampling points are first selected to establish the prior distribution of the surrogate model. Subsequently, 80 iterations are performed based on the Gaussian process surrogate model, where the expected improvement (EI) criterion is employed to dynamically balance exploration and exploitation on a global scale, thereby gradually approaching the optimal set of conditions. To present the optimization results more intuitively, after completing the three-dimensional optimization, the objective function is projected onto two-dimensional grids for visualization, and heatmaps are plotted with the optimal points marked, as shown in Figure 9, Both BYD yield and HCHO conversion exhibited a consistent optimization trend under the coupled effects of time, temperature and pH. Among these factors, pH has the most significant influence, with both metrics reaching their maxima in the neutral to near-neutral range (pH: 7.1–7.3), showing a clear synergistic enhancement effect. These results are highly consistent with the experimental observations, further validating the reliability and industrial applicability of the constructed model and optimization framework.

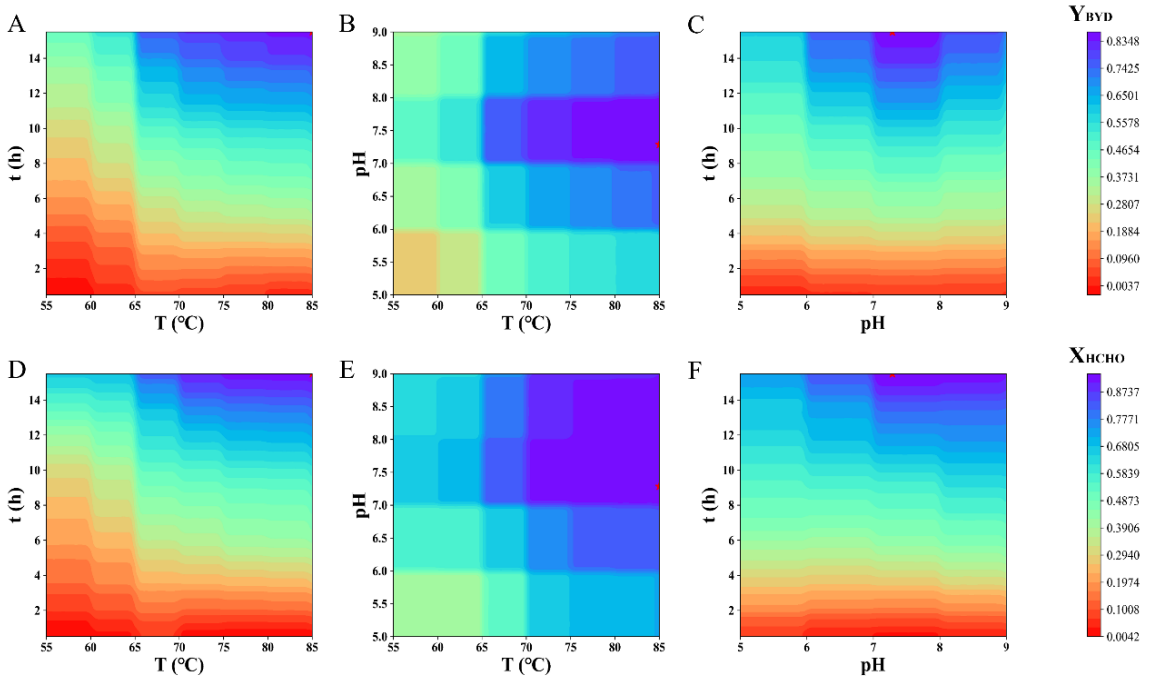


Figure 9. The heat map of the cross-influence of various influencing factors. (**A**–**C**) BYD yield; (**D**–**F**) HCHO conversion.

Overall, the proposed modeling strategy effectively combines physical rationality with data-driven approaches, providing valuable guidance for process optimization during the synthesis of 1,4-butanediol in slurry

bed reactors and offering a transferable pathway for small-sample modeling and condition optimization in other complex reaction systems. Table 3 presents a qualitative comparison between this study and recent machine learning-based kinetic modeling works, highlighting the advantage of our approach with a significantly higher R^2 value, demonstrating improved accuracy and robustness. While the MAE in our study is slightly higher compared to some models, the substantially higher R^2 value reinforces the superior predictive capability and generalization ability of our model, particularly in scenarios with limited experimental data. This further emphasizes the reliability and stability of our model in real-world applications.

Table 3. Comparison of model performance with previous studies.

Study	Reaction Type	Modeling Framework	R^2	MAE/RMSE
VOC concentration prediction (Ref. [35])	VOC concentration control	RF, SVM, XGBoost	=0.93 (max)	
Bayesian ML model (Ref. [51])	Catalytic O_3 degradation	CatBoost + Bayesian Optimization	0.9482	RMSE = 0.0401
Ensemble ML (Ref. [35])	mRNA–lipid nanoparticle formulation	XGBoost + Bayesian Optimization + Ensemble (SVEM)	>0.94	N/A
This work	Formaldehyde–acetylene (BYD synthesis)	XGBoost + Data Augmentation + Physical Constraints + Bayesian Optimization	0.982	MAE = 0.058

5. Conclusions

This work focused on the synthesis of 1,4-butynediol in a slurry bed reactor, especially investigated the essential role of pH in determining reaction selectivity and yield. Based on limited experimental data, a physics-constrained data augmentation strategy was introduced and integrated with machine learning model to construct a data-driven model. Four representative algorithms, i.e., RF, ET, LightGBM, and XGBoost were compared in terms of predictive accuracy, generalization, and stability. The optimal model (XGBoost) was coupled with Bayesian optimization to perform dual-objective optimization targeting both yield and conversion. The main conclusions are summarized as follows:

(1) Through physics-constrained interpolation, noise perturbation, and conservation filtering, the original 84 experimental data can be expanded into 1023 valid samples, which enabled smooth kinetics and chemically consistent trends.

(2) Model comparison showed XGBoost outperformed RF, ET, and LightGBM, achieving the highest predictive accuracy ($R^2 = 0.9847 \pm 0.0022$ and MAE $= 0.0204 \pm 0.0010$ for predicting HCHO conversion, $R^2 = 0.9773 \pm 0.0035$ and MAE $= 0.0207 \pm 0.0012$ for predicting BYD yield).

(3) Operating under near-neutral conditions (pH 7.1–7.3) leads to the simultaneous maximization of BYD yield and HCHO conversion. This trend was consistently validated through global Bayesian multi-objective optimization, underscoring the critical regulatory role of pH in the reaction system.

Although the proposed model exhibits high predictive accuracy, it is limited to laboratory-scale datasets and does not consider catalyst deactivation or large-scale heat/mass transfer effects. Future research will integrate real-time monitoring, adaptive ML control, and catalyst stability modeling to achieve continuous optimization in industrial SBR systems.

Supplementary Materials

The additional data and information can be downloaded at: https://media.sciltp.com/articles/others/25120509162683 84/SCE-25080344-Supplementary-Materials-FC-done.pdf. Figure S1: Standard curve of propynol concentration. Figure S2. Standard curve of 1,4-butynyldiol concentration. Figure S3: Parallel experiment error graph. Table S1: Raw Data of Figure 3A to 3C. Table S2: Raw Data of Figure 3D to 3F. Table S3: Raw Data of Figure 4 and Figure 5A,B. Table S4: Raw Data of Figure 6A. Table S5: Raw Data of Figure 6B. Table S6: Raw Data of Figure 7A. Table S7: Raw Data of Figure 7B. Table S8: Raw Data of Figure 7C. Table S9: Raw Data of Figure 7D. Table S10: Raw Data of Figure 8A. Table S11: Raw Data of Figure 8B.

Author Contributions

X.-Q.L.: conceptualization, methodology, software, data curation, writing—original draft preparation; Z.-Q.J.: visualization, investigation; H.-L.W.: writing—reviewing and editing, supervision, software, validation; Z.-H.L.: writing—reviewing and editing, supervision, validation; All authors have read and agreed to the published version of the manuscript.

Funding

The authors thank the Joint Funds of the National Natural Science Foundation of China (No. U24A20529) and the key R & D project of Ningxia Autonomous Region (No. 2024BEE02004) for supporting this work.

Data Availability Statement

All of the experimental and numerical data will be available on request.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used DeepSeek to polish the language and improve the fluency of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Nomenclature/Abbreviation

Nomenclature

C^0_{HCHO}	the concentration of HCHO at the initial moment, [mol/L]
C^t_{HCHO}	the concentration of HCHO at time *t*, [mol/L]
C^t_{BYD}	the concentration of BYD at time *t*, [mol/L]
T	temperature, [℃]
t	time, [h]
X	the conversion of HCHO, [-]
Y	the yiled of BYD, [-]
R^2	coefficient of determination, [-]
MAE	mean absolute error, [-]
$\boldsymbol{Y(\varphi)}$	objective function, [-]

Abbreviation

BDO	1,4-butanediol
BYD	2-butyne-1,4-dioy
PA	propargyl alcohol
HCHO	formaldehyde
PBT	poly butylene terephthalate
GBL	γ-butyrolactone
PU	polyurethane
NMP	N-methylpyrrolidone
PBS	poly butylene succinate
PLC	programmable logic controller
SBR	slurry bed reactor
GC	gas chromatography
GC-MS	gas chromatography—mass spectrometry
FID	flame ionization detector
RF	random forest
ET	extremely randomized trees
LightGBM	light gradient boosting machine
XGBoost	extreme gradient boosting
GP	gaussian process
SVM	support vector machine
EOSs	equations of state
VOC	volatile organic compounds
SLM	selective laser melting
BO-GPR	bayesian optimization based on gaussian process regression
EI	expected improvement

References

1. Trotuş, I.; Zimmermann, T.; Schüth, F. Catalytic Reactions of Acetylene: A Feedstock for the Chemical Industry Revisited. *Chem. Rev.* **2014**, *114*, 1761–1782.
2. Chen, X.; Zhang, M.; Yang, K.; et al. Raney Ni–Si Catalysts for Selective Hydrogenation of Highly Concentrated 2-Butyne-1,4-diol to 2-Butene-1,4-diol. *Catal. Lett.* **2014**, *144*, 1118–1126.
3. Zawadzki, B.; Abid, R.; Fernandez-Ropero, A.J.; et al. Effect of Iron Oxidation State on the Catalytic Performance of Fe/C in Liquid Phase Flow Hydrogenation of 2-Butyne-1,4-diol. *Fuel* **2025**, *380*, 133170.

4. D'Amboise, M.; Mathieu, D.; Piron, D.L. A Chemical Study of 2-Butyne-1,4-diol. *Talanta* **1988**, *35*, 763–768.
5. Hosseini, M.G.; Arshadi, M.R. Study of 2-Butyne-1,4-diol as Acid Corrosion Inhibitor for Mild Steel with Electrochemical, Infrared and AFM Techniques. *Int. J. Electrochem. Sci.* **2009**, *4*, 1339–1350.
6. Yang, W.; Peng, W.; Li, H.; et al. Catalytic Ethynylation of Formaldehyde for Selective Propargyl Alcohol Production Using the Copper Metal Organic Framework HKUST-1. *New J. Chem.* **2024**, *48*, 9082–9089.
7. Tanielyan, S.K.; More, S.R.; Augustine, R.L.; et al. Continuous Liquid-Phase Hydrogenation of 1,4-Butynediol to High-Purity 1,4-Butanediol over Particulate Raney Nickel Catalyst in a Fixed Bed Reactor. *Org. Process Res. Dev.* **2017**, *21*, 327–335.
8. Yang, G.; Yang, L.; Chen, J. Effective Performance of the Cu/Zn/SiO_2 Catalyst Applied in the Ethynylation of Formaldehyde for 1,4-Butynediol Synthesis. *Ind. Eng. Chem. Res.* **2023**, *62*, 21067–21077.
9. Wang, C.; Hai, X.; Bai, J.; et al. Elucidating the Atomic Stacking Structure of Nickel Phyllosilicate Catalysts and Their Consequences on Efficient Hydrogenation of 1,4-Butynediol to 1,4-Butanediol. *Chem. Eng. J.* **2024**, *488*, 150723.
10. Wang, Z.; Ban, L.; Meng, P.; et al. Ethynylation of Formaldehyde over CuO/SiO_2 Catalysts Modified by Mg Species: Effects of the Existential States of Mg Species. *Nanomaterials* **2019**, *9*, 1137.
11. Franz, A.W.; Kircher, M. Options for CO_2-Neutral Production of Bulk Chemicals. *J. Bus. Chem.* **2021**, *18*, 63–78.
12. Yang, G.; Yu, Y.; Tahir, M.U.; et al. Promotion Effect of Bi Species in Cu/Bi/MCM-41 Catalysts for 1,4-Butynediol Synthesis by Ethynylation of Formaldehyde, Reaction Kinetics. *React. Kinet. Mech. Catal.* **2019**, *127*, 425–436.
13. Wang, D.; Li, Y.; Yang, Y.; et al. Process Reconfiguration for the Production of 1, 4-Butanediol Integrating Coal with Off-Grid Renewable Electricity. *Int. J. Hydrogen Energy* **2025**, *102*, 1295–1305.
14. Yang, G.; Gao, F.; Yang, L. The Importance of Copper-Phyllosilicate Formed in CuO/SiO_2 Catalysts in the Ethynylation of Formaldehyde for 1,4-Butynediol Synthesis. *React. Chem. Eng.* **2023**, *8*, 881–890.
15. Zhao, F.; Ikushima, Y.; Arai, M. Hydrogenation of 2-Butyne-1,4-diol in Supercritical Carbon Dioxide Promoted by Stainless Steel Reactor Wall. *Catal. Today* **2004**, *93*, 439–443.
16. Yeston, J.; Coontz, R. Chemistry Writ Large. *Science* **2009**, *325*, 691.
17. Li, L.; Wei, X.; Lv, S.; et al. Specific Catalytic Hydrogenation of 2-Butyne-1,4-diol to Butane-1,4-diol. *Fuel* **2025**, *396*, 134673.
18. Rode, C.V.; Tayade, P.R.; Nadgeri, J.M.; et al. Continuous Hydrogenation of 2-Butyne-1,4-diol to 2-Butene- and Butane-1,4-diols. *Org. Process Res. Dev.* **2006**, *10*, 278–284.
19. Wei, H.L.; Liu, X.Q.; Jihou, Z.Q.; et al. Synthesis of But-2-yne-1, 4-diol in a Slurry Bed Reactor: Mechanisms, Kinetics and Process Optimization. *Chem. Eng. Sci.* **2025**, *320*, 122665.
20. Prats, H.; Illas, F.; Sayós, R. General Concepts, Assumptions, Drawbacks, and Misuses in Kinetic Monte Carlo and Microkinetic Modeling Simulations Applied to Computational Heterogeneous Catalysis. *Int. J. Quantum Chem.* **2017**, *118*, e25518.
21. Zhou, X.; Zhang, J.; Zhang, M. Active Site Reconstruction of a Metal Hydroxide/Metal Molybdate Heterogeneous Interface Enhances Electrochemical Water Oxidation. *Inorg. Chem. Front.* **2025**, *19*, 5819–5829.
22. Carbonaro, N.J.; Thorpe, I.F. Using Structural Kinetic Modeling to Identify Key Determinants of Stability in Reaction Networks. *J. Phys. Chem. A* **2017**, *121*, 4982–4992.
23. Wei, H.L.; Ma, X.M.; Qin, J.Z.; et al. Image-Based Method for In-Situ Monitoring of Reaction Kinetics. *Chem. Eng. Sci.* **2025**, *320*, 122720.
24. Jin, J.; Ni, L.; Qiu, W.; et al. Kinetic Evaluation for the Reaction of Hydroxylamine with Acetamide Using Online Infrared Spectra and pH Profile Analysis, Reaction Kinetics. *React. Kinet. Mech. Catal.* **2023**, *136*, 1819–1837.
25. Milani, G.; Milani, F. Parabola-Hyperbola pH Kinetic Model for NR Sulphur Vulcanization. *Polym. Test.* **2017**, *58*, 104–115.
26. Blurock, E.S. Characterizing Complex Reaction Mechanisms Using Machine Learning Clustering Techniques. *Int. J. Chem. Kinet.* **2004**, *36*, 107–118.
27. Kayala, M.A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52*, 2526–2540.
28. Dias, L.S.; Ierapetritou, M.G. Integration of Planning, Scheduling and Control Problems Using Data-Driven Feasibility Analysis and Surrogate Models. *Comput. Chem. Eng.* **2019**, *134*, 106174.
29. Regis, R.G. Multi-Objective Constrained Black-Box Optimization Using Radial Basis Function Surrogates. *J. Comput. Sci.-Neth.* **2016**, *16*, 140–155.
30. Marino, D.L.; Manic, M. Physics Enhanced Data-Driven Models with Variational Gaussian Processes. *Ieee Open J. Ind. Elec.* **2021**, *2*, 252–265.
31. Xia, Y.; Dai, L.; Xie, W.; et al. Network-Based Data-Driven Filtering with Bounded Noises and Packet Dropouts. *Ieee T. Ind. Electron.* **2016**, *64*, 4257–4265.
32. Ye, H.; Du, Z.; Lu, H.; et al. Using Machine Learning Methods to Predict VOC Emissions in Chemical Production with Hourly Process Parameters. *J. Clean. Prod.* **2022**, *369*, 133406.

33. Mohammadi, M.; Hadavimoghaddam, F.; Atashrouz, S.; et al. Modeling Hydrogen Solubility in Alcohols Using Machine Learning Models and Equations of State. *J. Mol. Liq.* **2022**, *346*, 117807.
34. Qi, G.; Liu, B. Production Feature Analysis of Global Onshore Carbonate Oil Reservoirs Based on XGBoost Classier. *Processes* **2024**, *12*, 1137.
35. Maharjan, R.; Kim, K.H.; Lee, K.; et al. Machine Learning-Driven Optimization of mRNA-Lipid Nanoparticle Vaccine Quality with XGBoost/Bayesian Method and Ensemble Model Approaches. *J. Pharm. Anal.* **2024**, *14*, 100996.
36. Xiang, Y.; Pan, B.; Luo, L. A New Model Updating Strategy with Physics-Based and Data-Driven Models. *Struct. Multidiscip. Optim.* **2021**, *64*, 163–176.
37. Jiang, J.; Zhang, C.; Ke, L.; et al. A Review of Machine Learning Methods for Imbalanced Data Challenges in Chemistry. *Chem. Sci.* **2025**, *16*, 7637–7658.
38. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90.
39. Zhang, X.; Chen, B.; Wang, J.; et al. Review of Molybdenum Disulfide Research in Slurry Bed Heavy Oil Hydrogenation. *ACS Omega* **2023**, *8*, 18400–18407.
40. Gu, P.; Zhang, Y.; Du, H. Experimental Study on Back-Flushing Characteristics of an In-Vessel Filtration System in Fischer–Tropsch Slurry Reactors. *Ind. Eng. Chem. Res.* **2023**, *62*, 17937–17946.
41. Heracleous, E.; Papadopoulou, F.; Lappas, A.A. Continuous Slurry Hydrotreating of Sewage Sludge-Derived Hydrothermal Liquefaction Biocrude on Pilot-Scale: Comparison with Fixed-Bed Reactor Operation. *Fuel Process. Technol.* **2024**, *253*, 108006.
42. Cubuk, E.D.; Zoph, B.; Shlens, J.; et al. *Randaugment: Practical Automated Data Augmentation with a Reduced Search Space*, 3rd ed.; Computer Science Press: Beijing, China, 2019; pp. 702–703.
43. Liu, F.; Chen, H.; Yang, J.; et al. Application of Physics-Informed Machine Learning Methods in Buckling Design of Axially Compressed Cylindrical Shells. *Thin Wall Struct.* **2024**, *200*, 11963.
44. Zhang, X.; Gong, J.; Xuan, F. A Physics-Informed Neural Network for Creep-Fatigue Life Prediction of Components at Elevated Temperatures. *Eng. Fract. Mech.* **2021**, *258*, 108131.
45. Karpatne, A.; Atluri, G.; Faghmous, J.H.; et al. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318–2331.
46. Wang, H.; Li, B.; Xuan, F. A Dimensionally Augmented and Physics-Informed Machine Learning for Quality Prediction of Additively Manufactured High-Entropy Alloy. *J. Mater. Process. Tech.* **2022**, *307*, 117637.
47. Yu, A.; Pan, Y.; Wan, F.; et al. Rapid Accomplishment of Cost-Effective and Macro-Defect-Free LPBF-Processed Ti Parts Based on Deep Data Augmentation. *J. Manuf. Process.* **2024**, *120*, 1023–1034.
48. Isobe, H.; Xiao, X.; Fukunaga, T.M.; et al. Revealing Kinetic Features of a Macrocyclization Reaction Using Machine-Learning-Augmented Data. *Angew. Chem. Int. Ed.* **2025**, *137*, e202501365.
49. Hribar, U.; Stevanoska, S.; Camacho-Villalón, C.L.; et al. Optimizing Foamed Glass Production with Machine Learning. *Mater.Des* **2025**, *257*, 114459.
50. Zhang, F.; Tamura, R.; Zeng, F.; et al. Bayesian Optimization for Controlled Chemical Vapor Deposition Growth of WS_2. *Acs Appl. Mater. Inter.* **2024**, *16*, 59109–59115.
51. Wang, X.; Zheng, X.; Huang, Z.; et al. Prediction and Optimization of Key Factors for Catalytic O_3 Degradation of Antibiotics Based on Catboost Model Coupled Bayesian Optimisation Algorithm. *J. Water Process Eng.* **2025**, *72*, 107481.