Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

UNIVERSAL WISER PUBLISHER

Research Article

Data-Driven Optimization of C_4 Olefin Synthesis: A Gradient Boosting and Particle Swarm Optimization Framework

Qian Zhang[1†], Feng Wang[1†], Shimei Zhang[1], Lingling Luo[1], Zhuoyong Shi[2*]

[1] Department of Mathematics, Xi'an Jiaotong University City College, Xi'an, China
[2] Faculty of Science, National University of Singapore, Singapore

[†] These authors contributed equally to this work
E-mail: shizhuoyong@u.nus.edu

Received: 8 May 2025; **Revised:** 16 June 2025; **Accepted:** 24 June 2025

Abstract: The synthesis of C_4 olefins from ethanol presents a promising sustainable pathway, yet its industrial application is hindered by the challenge of optimizing complex, nonlinear interactions among reaction parameters. To address this, we developed a hybrid machine learning framework combining a Gradient Boosting Decision Tree (GBDT) for predictive modeling with Particle Swarm Optimization (PSO) for parameter optimization. The integrated GBDT-PSO model accurately captured the process dynamics, achieving a prediction accuracy of 92.5%. Our results revealed a critical reaction temperature threshold of 350 °C, above which C_4 olefin yields increase significantly. Optimal conditions were identified at 400 °C with a 2 wt% Co/SiO_2 + Hydroxyapatite (HAP) catalyst, achieving a maximum yield of 44.73%. The proposed framework markedly outperformed conventional optimization methods, including Genetic Algorithm (88.2%) and Differential Evolution (86.7%). This study not only provides a robust, data-driven methodology for process optimization but also delivers actionable insights that can guide the scale-up and industrial implementation of ethanol-to-olefin technologies.

***Keywords*:** ethanol conversion, C_4 olefins, gradient boosting tree, particle swarm optimization, catalyst optimization

MSC: 65K05, 90C59, 68T05, 80A32

Abbreviation

GBDT	Gradient Boosting Decision Tree
PSO	Particle Swarm Optimization
GXGB	Gradient Boosting Tree
SSA	Sparrow Search Algorithm
GBT	Gradient Boosting Tree
GA	Genetic Algorithm
DE	Differential Evolution
RF	Random Forest

DOI: https://doi.org/10.37256/cm.6420257205

1. Introduction

C_4 olefins are essential chemical intermediates with broad applications in the production of pharmaceutical chemicals and industrial materials [1]. Traditionally, C_4 olefins are primarily derived from catalytic cracking of heavy oil in refineries, naphtha cracking in ethylene plants, and as by-products of coal-to-olefin processes [2]. However, these approaches are resource intensive and environmentally detrimental, leading to increased interest in sustainable and renewable production methods. Among these, ethanol, as a renewable and environmentally friendly feedstock, has received significant attention for its potential in C_4 olefin synthesis [3].The use of ethanol for C_4 olefin production offers a novel approach to accessing C_4 hydrocarbon resources. Since 2015, ethanol has been efficiently converted into C_4 olefins via dehydration and coupling reactions. This method not only expands the potential applications of biomass ethanol but also establishes a sustainable pathway for C_4 olefin production. Despite these advantages, the conversion process is heavily influenced by factors such as catalyst combinations and reaction temperature. Optimizing these parameters to enhance the yield of C_4 olefins remains an urgent and complex challenge.

To solve this problem, one part of researchers used several theoretical approaches and optimized the synthesis of C_4 olefins. In 2022, Fang et al. [4] investigated the effect of catalyst combination and temperature on the C_4 olefin yield using multiple linear regression and time series prediction models, which provided important information to understand the reaction mechanism and optimizing the reaction conditions. In 2023, Li et al. [5] explored the optimal catalyst combination by using Decision Tree Modeling. In 2022, Zhang et al. [6] optimized the process conditions for the preparation of C_4 olefins by ethanol catalytic coupling through machine learning and multivariate non-linear fitting methods, and achieved certain results. In 2023, Huang et al. [7] established a model based on a generalized transverse Algorithm to predict the optimal combination of ethanol conversion ratio and C_4 olefin production. In 2022, Zhang et al. [8] explored the optimal catalyst and temperature conditions for the production of C_4 olefins by catalytic coupling of ethanol, which provided an important guideline for obtaining the highest yield. In 2023, Zhou et al. [9] proposed a hybrid model based on Sample incremental eXtreme Gradient Boosting Tree and Sparrow Search Algorithm for optimizing the production conditions of C_4 olefins. In 2022, Bi's [10] research on nano HZSM-5 molecular sieve catalysts showed that the use of such catalysts can effectively reduce the reaction temperature and simplify the process flow, thus significantly improving the energy efficiency of the process. In 2018, Lv [11] discussed the preparation of butanol and C_4 olefins by ethanol coupling and proposed an optimization model to improve the yield and selectivity.

All of these people [12–14] have studied gradient boosted trees and proposed variants of the original algorithm to improve the training speed in turn. Didrik [15] argue that Gradient tree boosting is a powerful machine learning technique that has shown good performance in predicting a variety of outcomes. Wang et al. [16] studyed that Gradient tree boosting is a prediction algorithm that sequentially produces a model in the form of linear combinations of decision trees, by solving an infinite-dimensional optimization problem. Nakhaei-Kohani et al. [17] explored the GBDT algorithm and found that GBDT excelled in prediction accuracy by comparing it with other machine learning models such as Random Forest, Decision Tree, etc. Kou et al. [18] describes gradient boosting trees to optimize the model by combining multiple weak learners in an iterative manner, embodying the strategy of reaching a globally optimal solution through gradual improvement. Hatwell et al. [19] proposed the gbt-HIPS method, a novel heuristic algorithm for interpreting decision rules in GBT classification models. Adler [20] focuses on the issue of Feature Importance (FI) with GBT. Zhou et al. [21] explored a new tree boosting variable coefficient model framework proposed by combining the flexibility of gradient boosting trees, which provides new perspectives and possibilities for the application of gradient boosting trees in the field of statistical modeling. Biau et al. [22] is of great significance for improving the computational efficiency and interpretability of machine learning models. Luo, Freund and Friedman et al. [23–25] introduced the Gradient Boosting Tree algorithm, which demonstrated the efficiency and accuracy of the Gradient Boosting Tree in dealing with multi-class classification tasks. A solid theoretical foundation is laid for the development and application of the subsequent gradient boosting tree algorithm.

Ardiansyah et al. [26] propose an improved PSO algorithm, which improves the convergence speed and diversity through pseudo-random sequences and opposite-rank inertia weights, and the experimental results outperform Genetic Algorithm (GA) and Improved Genetic Algorithm (IGA). The PSO algorithm has become an efficient search method for solving complex optimization problems since it was proposed by Valdez [27] in 2020. Clerc and Kennedy [28] researchers have explored the dynamic properties of PSO in depth, and have proposed several improved variants to enhance its performance. Schlauwitz et al. [29] compared the particle swarm optimization algorithm with the traditional particle swarm optimization algorithm and differential evolution algorithm, and the experimental results showed that the algorithm performs better on high-dimensional problems. The PSO algorithm [30] has become a popular heuristic algorithm for solving complex optimization problems since it was proposed in 1995, and its property of mimicking the intelligent behavior of groups of animals has led to its wide application in several fields. PSO technique [27] based on Swarm algorithm effectively solves optimization problems in chemometrics by simulating collective behavior of social animals.

Despite the results of the above studies, there are still some shortcomings. Some of the studies still have limitations in dealing with complex nonlinear relationships and multivariate optimization, ignoring the complex interactions among different variables, resulting in insufficient prediction accuracy and applicability of the models.

In order to improve the above deficiencies, this paper proposes an optimization design method for C_4 olefin synthesis process based on gradient boosting tree and particle swarm algorithm. By analyzing and processing the experimental data, we obtained a number of factors that have a greater impact on the C_4 olefin yield, and then established the GBDT model. The general PSO algorithm is used to optimize and solve the GBDT model to obtain the optimal oleffn yield under certain conditions, which provides new ideas for the optimization of complex reaction processes in the chemical industry.

2. Literature review

The synthesis of C_4 olefins from ethanol has been studied through multiple approaches.

2.1 *Catalyst optimization studies*

Bi et al. [10] demonstrated that nano HZSM-5 molecular sieve catalysts can reduce reaction temperatures by 15-20% compared to traditional catalysts. Subsequent work by Lv [11] further improved selectivity through cobalt-modified zeolites.

2.2 *Machine learning applications*

Recent advances include:

Zhang et al. [6] used multivariate nonlinear fitting to optimize yields.

Zhou et al. [9] combined XGBoost with Sparrow Search Algorithm.

2.3 *PSO in chemical engineering*

The PSO algorithm [27] has been widely adopted for parameter optimization:

$$v_i(t+1) = \omega v_i(t) + c_1 r_1 (p_i - x_i(t)) + c_2 r_2 (g - x_i(t)). \tag{1}$$

In Eq. (1), ω is the inertia weight (typically 0.4-0.9) [26].

3. Methods and analysis

3.1 *Study framework*

This research systematically combines theory, modeling, and validation to optimize C_4 olefin production in Figure 1. It begins with *literature review* (catalysts, machine learning), followed by *data preprocessing* (interpolation, correlation analysis). A *GBDT model* captures nonlinear relationships between reaction parameters (temperature, catalyst ratios), while *PSO* optimizes model parameters. Results validate temperature thresholds (350 °C) and optimal catalyst combinations (e.g., Co/SiO$_2$ = 1 : 100), linking computational insights to practical applications. The framework ensures a rigorous transition from theory to industrial-relevant findings.

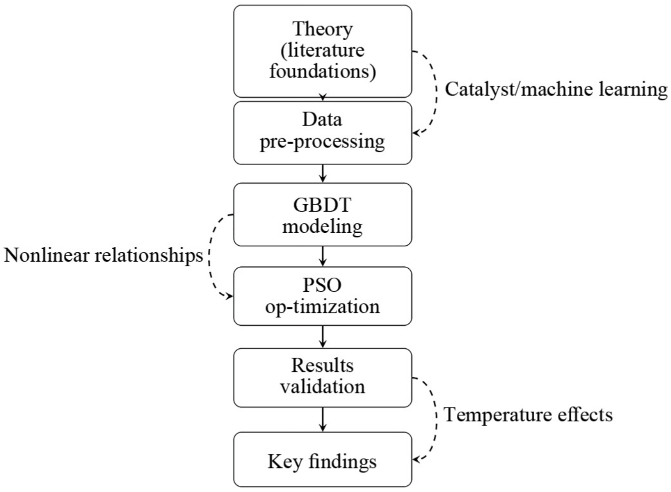

Figure 1. Study framework

3.2 *Data sources and preprocessing*

3.2.1 *Data sources*

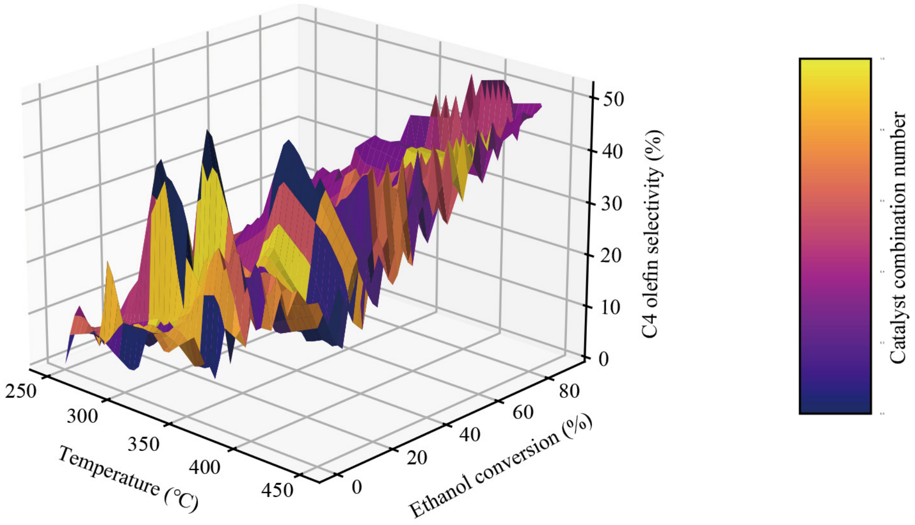

Figure 2. Trend graph of experimental data

In this paper, the data [5] were obtained from a number of ethanol conversion experiments carried out in a chemical laboratory with different catalysts at different temperatures, which recorded ethanol conversion and C$_4$ olefin selectivity at different catalyst combinations and reaction temperatures. The Figure 2 visualizes the overall trend of these experimental data in the form of a three-dimensional surface plot. However, due to the obvious gaps and discontinuities between the data points, the analysis results may be biased, which has a direct impact on the accuracy of the model we hope to establish in this paper. Therefore, the data needs to be further pre-processed so that each set temperature point can be analyzed with corresponding data.

3.2.2 *Data preprocessing*

In order to ensure that the conclusions of this paper have sufficient accuracy and stability, the Cubic Spline Interpolation Method is chosen to fill in the missing data. Cubic Spline Interpolation has high flexibility and accuracy, can fit the data smoothly, and is especially suitable for the completion of nonlinear relationships. According to the analysis of the experimental data, it was found that the ethanol conversion rate and C_4 olefin selectivity were in accordance with the logarithmic normal distribution, and the p-values of the Kolmogorov-Smirnov test were 0.42762090 and 0.80044529, respectively, which did not significantly deviate from the logarithmic normal distribution ($p > 0.05$) suggesting that the data complied with the assumption of logarithmic normal distribution. Therefore, the use of cubic spline interpolation can reflect the actual trend of the data more scientifically.

We know the cubic spline interpolation method with the following formula:

$$S_j(y) = a_j + b_j(y - y_j) + c_j(y - y_j)^2 + d_j(y - y_j)^3, \tag{2}$$

In Eq. (2), $j = 0,\ 1,\ 2,\ \ldots,\ n-1$. Which has n intervals and $n+1$ nodes. And the first and second order derivatives of $S_j(y)$ are continuous at all data points. The second order derivative of $S_j(y)$ is zero at the start and end points of the data.

We interpolate the initial data with cubic spline to obtain a 3D surface map as shown in Figure 3.

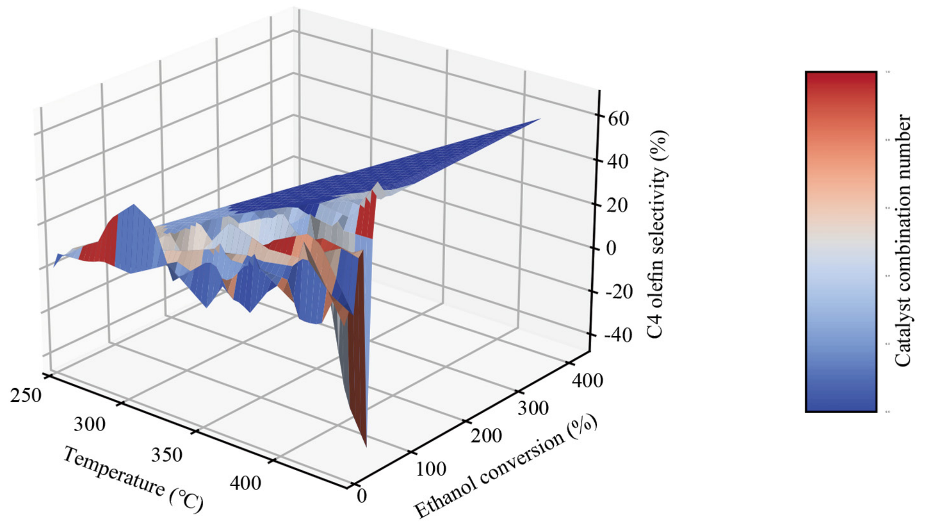

Figure 3. Three-dimensional surface map

After triple spline interpolation of the initial data, then we fit the processed data at this point to the temperature, as shown in Figure 4, where the ethanol conversion versus temperature fit is shown in Figure 4a and the C_4 olefin selectivity versus temperature fit is shown in Figure 4b.

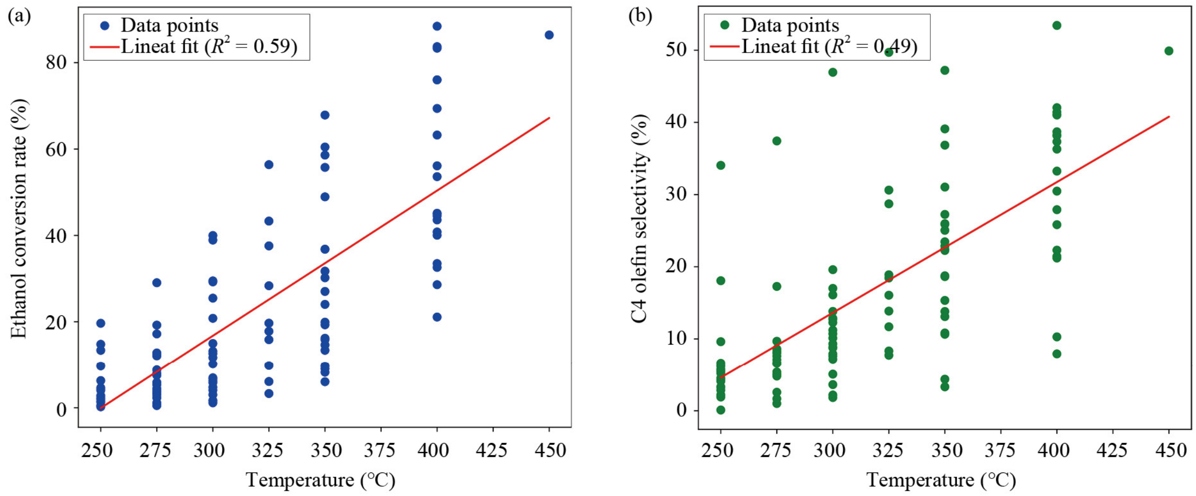

Figure 4. Temperature versus data fit. (a) Ethanol conversion fitted to temperature; (b) C_4 olefin selectivity with temperature fitting

The ethanol conversion and C_4 olefin selectivity generally increased with increasing temperatures in the range of 250 °C to 400 °C, as analyzed in Figure 4. The goodness of fit R^2 value is 0.59, indicating a moderate linear relationship between temperature and ethanol conversion ratio, and the linear fit showed a positive correlation between ethanol conversion ratio and temperature, as shown in Figure 4a. The linear fitting results show some linear relationship between C_4 olefin selectivity and temperature, but the correlation is weak, as described in Figure 4b. The overall trend suggests that an increase in temperatures contributes to improved ethanol conversion and C_4 olefin selectivity. However, the individual catalyst combinations will vary and require further analysis.

3.2.3 *Correlation analysis*

In this paper, the data were processed with the aid of Python development tools to obtain a partial table of the relationship between ethanol conversion and C_4 olefin selectivity for different catalyst combinations and temperatures (Table 1).

Table 1. Relationship between ethanol conversion and C_4 olefin selectivity at different catalyst combinations and temperatures (partial)

Catalyst combination number	Temperature	Ethanol conversion rate (%)	C_4 olefin selectivity (%)	Sio_2 mass	Co/Sio_2 mass	Co mass	HAP mass	Ethanol concentration	Quartz sand content
A1	250	2.06716	34.05	198	200	2	200	1.68	0
A1	275	5.85172	37.43	198	200	2	200	1.68	0
A1	300	14.968	46.94	198	200	2	200	1.68	0
A1	325	19.681	49.7	198	200	2	200	1.68	0
…	…	…	…	…	…	…	…	…	…
A11	250	0.22769	0.1	49.5	50	0.5	0	1.68	1
A11	275	0.51631	1	49.5	50	0.5	0	1.68	1
A11	300	1.6075	1.82	49.5	50	0.5	0	1.68	1
…	…	…	…	…	…	…	…	…	…
B7	300	11.7	12.86	99	100	1	100	0.9	0
B7	325	17.8	18.45	99	100	1	100	0.9	0
B7	350	30.2	25.05	99	100	1	100	0.9	0
B7	400	69.4	38.17	99	100	1	100	0.9	0

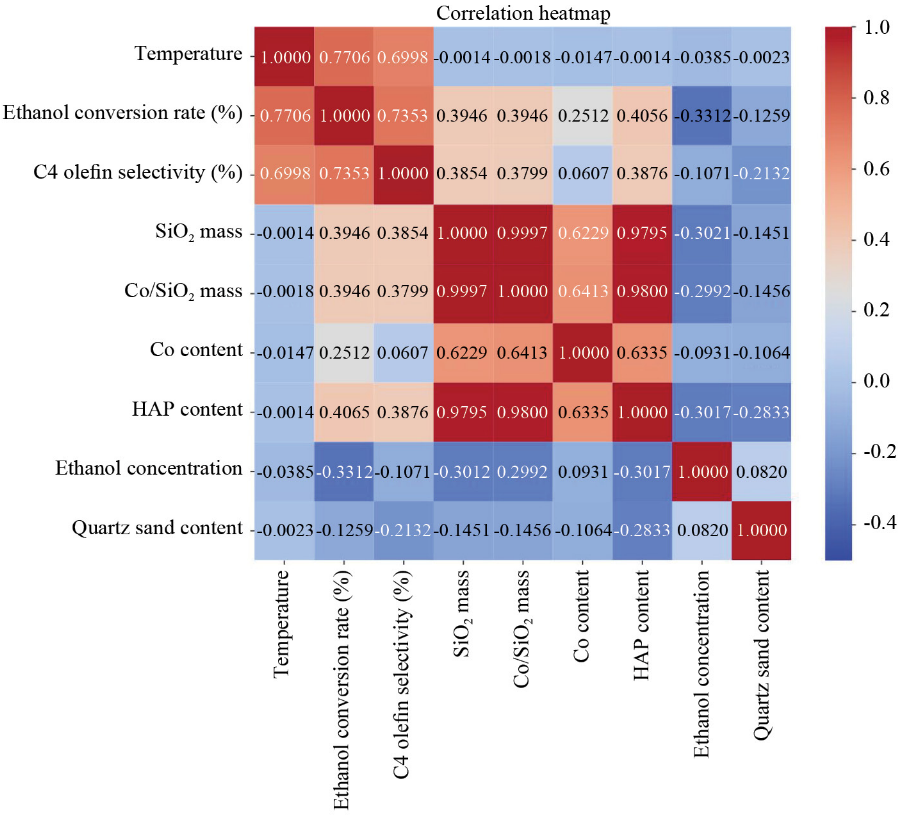

Figure 5. Correlation coefficient between variables

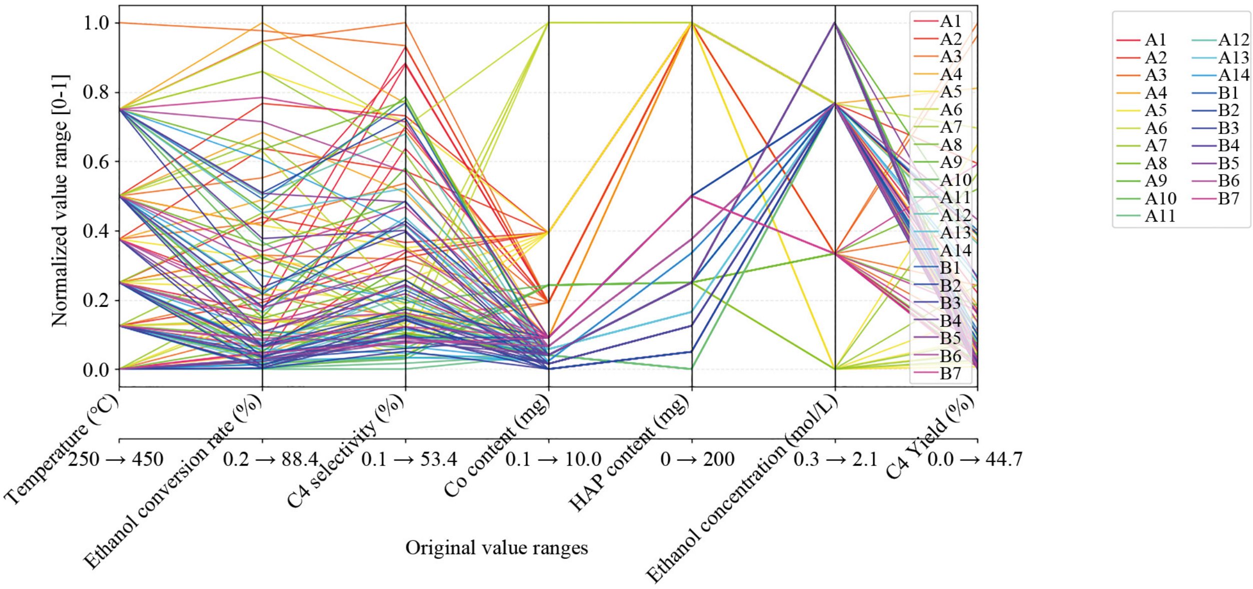

Figure 6. Catalyst performance analysis-parallel coordinates plot

Correlation analysis are performed on Table 1. A matrix of correlation coefficients between all the numerical variables in the data set is obtained, and the correlation coefficient diagram between the variables is obtained as shown in Figure 5.

The correlation coefficient plots not only reveal the interactions among the variables, but also screen out the key variables affecting ethanol conversion and C_4 olefin selectivity, i.e., the mass ratio of SiO_2, the mass ratio of Co/SiO_2, the Co content, the HAP content, the ethanol concentration, the quartz sand content, and the reaction temperature. Catalyst Performance Analysis-Parallel Coordinates Plot is utilized to analyze the performance of various catalysts across multiple performance indicators. The horizontal axis represents seven variables, namely temperature, ethanol conversion rate, C_4 selectivity, CO content, HAP content, ethanol concentration, and C_4 yield. The vertical axis indicates the normalized value range (0-1) in Figure 6. Each colored line corresponds to the performance data of a specific catalyst, with the intersections and distribution of the lines reflecting the relative performance across different indicators. Data normalization allows for comparison on a uniform scale, thereby facilitating the identification of catalyst strengths and weaknesses for specific indicators and providing an intuitive reference for catalyst selection and optimization.

3.2.4 *Quantitative analysis*

The C_4 olefin generation rate can be limited by the characteristics of the catalyst and the reaction environment. Among the many factors influencing this study, the catalyst is one of the key factors. The yield of C_4 olefin can be significantly improved by an efficient catalyst combination and a precise reaction environment. The definition of the the C_4 olefin yield as following equation:

$$Y_{C_4} = X_{\mathrm{ETOH}} \times S_{C_4}, \tag{3}$$

The C_4 olefin yield (Y_{C_4}) is calculated from the ethanol conversion rate (X_{ETOH}) and the C_4 olefin selectivity (S_{C_4}) in Eq. (3).

In order to further explore how different catalyst combinations and reaction conditions affect the conversion of ethanol and the selectivity of C_4 olefins, a multivariate correlation matrix thermogram was developed to reveal the interrelationships between the variables. The impact of each parameter can be quantified by analyzing the heat map, and in order to visualize the impact of these parameters, a Python-based radar diagram is created, as shown in Figure 7.

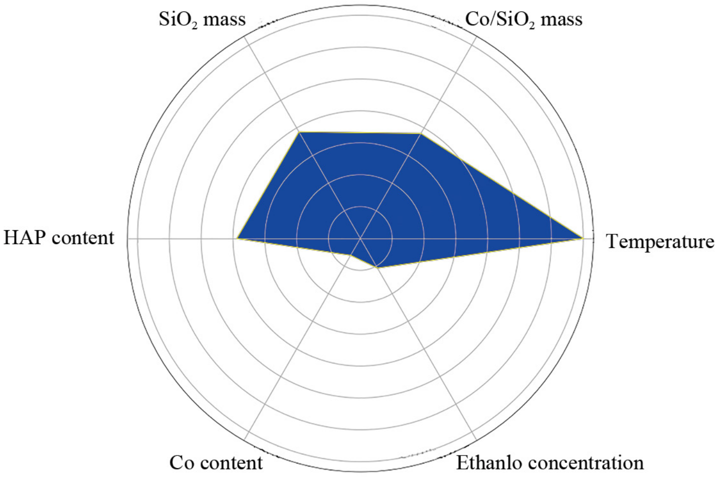

Figure 7. Influence of each parameter on the yield of C_4 olefins

It can be seen that the C_4 olefin yield is significantly affected by the reaction temperature, catalyst, SiO_2 quality and HAP content, see Figure 7. Among the many influencing variables, the temperature is obviously the most central one, which plays a decisive role in the conversion efficiency of ethanol and the selectivity of C_4 olefins.

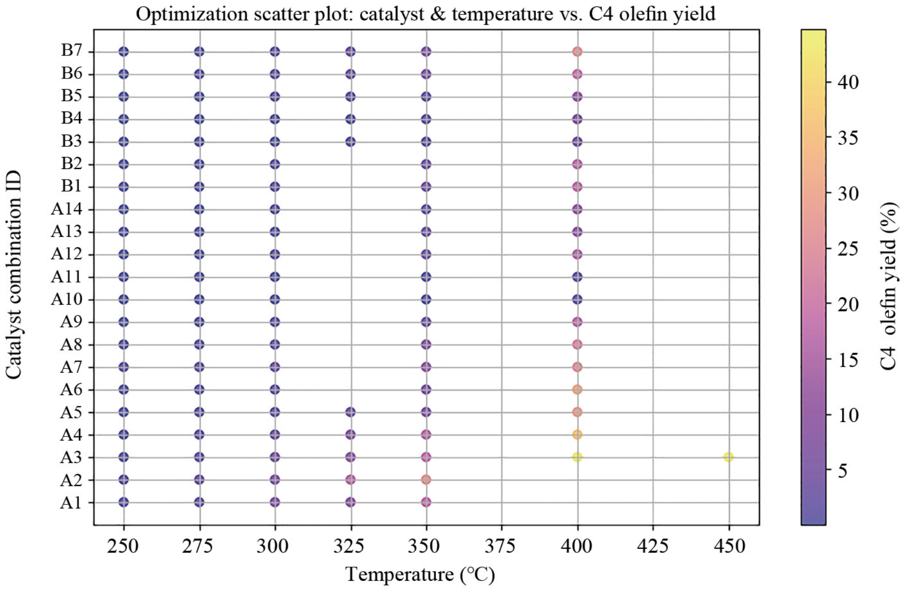

Figure 8. Conditional optimization scatterplot

However, the various combinations of the catalyst strategies have also led to significant fluctuations and variations in the C_4 olefin yields. For high activity and stability of the catalysts, it is necessary to rationally controlled to ensure that the yield of the final products reaches the desired value. Therefore, in order to gain insight into the specific effects of temperature and catalyst combination on the yield of C_4 olefins, a scatter plot of condition optimization is plotted with the help of Python, as shown in Figure 8. Each point in the graph symbolizes a specific experimental environment, and the shades of the color can reflect the relative differences in the yield of C_4 olefin.

It can be seen that the C_4 olefin yields of boyh A and B catalysts are higher when the temperature is higher than 350 °C, with the A3 catalyst combination showing the best performance at high temperatures, as shown in Figure 8. However, when the temperature is lower than 350 °C, the B catalysts are relatively ineffective, and the C_4 olefin yield of the A2 catalyst combination is the highest. These results provide a basis for subsequent catalyst selection.

3.3 *Model building*

In the conversion of ethanol to C_4 olefins, the catalyst combination and reaction temperature are the key factors affecting the yield. Due to the complex nonlinear relationship between reaction conditions and yield, a model capable of handling this complexity is needed. In order to achieve the optimization of the ethanol conversion process, GBDT model and PSO algorithm are chosen in this paper.

3.3.1 *Gradient boosting decision tree model*

The GBDT model is good at dealing with complex nonlinear relationships, especially in the chemical reaction process. The GBDT model accurately predicts the C_4 olefin yield under different conditions during the ethanol conversion process by constructing a series of decision trees to gradually approximate the optimal solution. The advantages of the GBDT model are described as following:

1) Dealing with nonlinear relationships: In the conversion of ethanol to C_4 olefins, there are complex nonlinear relationships between the catalyst combinations and reaction temperatures, which are difficult to be captured by traditional linear models. GBDT can effectively capture the nonlinear relationships and improve the prediction accuracy by accumulating multiple decision trees.

2) Efficient feature selection: GBDT can automatically select the features that have the greatest impact on the target variables. During the ethanol conversion process, the reaction yield under different catalysts and temperature conditions, and the GBDT model could identify the most important catalyst combinations and temperature parameters.

3) Powerful generalization: GBDT is excellent in handling high-dimensional data and preventing overfitting, and adapts well to noises and outliers in experimental data.

The expression of the GBDT model is as follows:

$$F_T(x) = F_0(x) + \nu \sum_{t=1}^{T} \sum_{j=1}^{J_t} \gamma_{jt} I(x \in R_{jt}), \tag{4}$$

In Eq. (4), $F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i,\ \gamma)$, L is the loss function, y_i is the true value of the sample, ν is the learning rate $(0 < \nu < 1)$ to control the impact of each trees on the final model, $\sum_{t=1}^{T} \sum_{j=1}^{J_t} \gamma_{jt}$ is the value γ_{jt} for each tree t (from 1 to T , where T is the number of trees), and for each leaf node j in each tree (from 1 to J_t , where J_t is the total number of leaf nodes in the tth tree), which is calculated to minimize the loss of the model. This value is the optimal step size calculated to minimize the loss of the model, and $I(x \in R_{jt})$ is an indicator function that indicates whether the sample x falls within the range of a particular leaf node R_{jt} in the decision tree.

Specifically, the steps for the application of the GBDT model in the conversion of ethanol to C_4 olefins are as follows:

1) Data pre-processing: Cleaning and standardization of the experimental data to ensure the data quality.

2) Characterization engineering: Extraction of key features that affect the conversion ratio of ethanol and the selectivity of C_4 olefins, such as catalyst type, reaction temperature, and so on.

3) Model training: Train the GBDT model with the preprocessed data, adjusting parameters such as the number of trees, depth, and learning rate.

4) Model Evaluation: Evaluate the predictive performance of the model through cross-validation and other methods to ensure that it has good generalization ability.

Through the above steps, the GBDT model can effectively solve the complex nonlinear relationship problems existing in the process of converting ethanol to C_4 olefins, and provide high-precision prediction results to optimize the catalyst combinations and reaction temperatures, so as to improve the yield of C_4 olefin. This method is not only theoretically important, but also provides effective technical support for the actual production process.

3.3.2 *Particle swarm optimization algorithm*

After constructing the GBDT model, the parameters of the model need to be further optimized to improve the prediction accuracy. PSO was selected for optimizing model parameters because of its excellent performance in dealing with complex nonlinear relationships and high-dimensional parameter spaces. Compared to other evolutionary algorithms, such as genetic algorithms and differential evolutionary algorithms, PSO typically has faster convergence and lower computational complexity, which makes it particularly suitable for parameter optimization in such chemical reaction processes.

The advantages of the PSO algorithm include the following:

1) Global search capability: The PSO algorithm is able to search globally in a wide range of parameter spaces to avoid falling into local optimal solutions. This is especially important for parameter optimization of GBDT model.

2) Fast convergence: The PSO algorithm has a fast convergence speed, which can find the optimal combinations of parameters in a relatively short period of time and improve the training efficiency of the model.

3) Easy to implement: The PSO algorithm is simple to understand, easy to implement and adjust, and is suitable for integration with GBDT models.

The specific steps of using the PSO algorithm are as follows:

1) Initialize the particle swarm: Randomly generate the initial particle swarm, each particle represents a GBDT model parameters set.

2) Fitness evaluation: Calculate the fitness value of each particle, i.e. the prediction error of the GBDT model under the corresponding parameter combinations. The lower the fitness value, the better the prediction performance of the model.

3) Update particle position and velocity: Update the velocity and position of each particle according to the individual optimal position and global optimal position, which are defined as follows:

$$v_i(t+l) = \omega v_i(t) + c_l r_l(p_i - x_i(t)) + c_2 r_2(g - x_i(t)), \tag{5}$$

4) Iterative optimization: Repeat the above steps until a preset number of iterations is reached or the fitness value converges.

3.3.3 *Comparative algorithms*

To verify the effectiveness of the GBDT-PSO model in optimizing the C_4 olefin synthesis process, we have selected the following commonly used optimization algorithms for comparison:

1) Genetic Algorithm (GA): An optimization algorithm based on the principles of natural selection and genetics that targets the objective function through cross-over, mutation, and selection operations.

2) Differential Evolution (DE): A population-based optimization algorithm that generates new candidate solutions through differential and mutation operations.

3) Random Forest (RF): An ensemble learning algorithm that improves prediction accuracy by constructing multiple decision trees.

These algorithms have been widely used in optimization fields, and comparisons can better demonstrate the advantages of the GBDT-PSO model.

3.3.4 *Comparative experiment design*

Comparative experiments were conducted using the same experimental dataset as the GBDT-PSO model, running GA, DE, and RF models respectively. The experimental setup is as follows:

1) Dataset Division: The experimental dataset was randomly divided into training and testing sets at a ratio of 80% training set to 20% testing set.

2) Parameter Settings: For each algorithm, Python is used to train according to the divided dataset to obtain the optimal parameters of each algorithm, and then the different parameters are compared.

3) Mean Squared Error (MSE) and Mean Absolute Error (MAE) were used as evaluation metrics to calculate the prediction accuracy of each algorithm in the test set.

4. Results

4.1 *GBDT modeling results*

Based on the constructed GBDT model, the feature importance relationships are plotted by PyCharm software as shown in Figure 9.

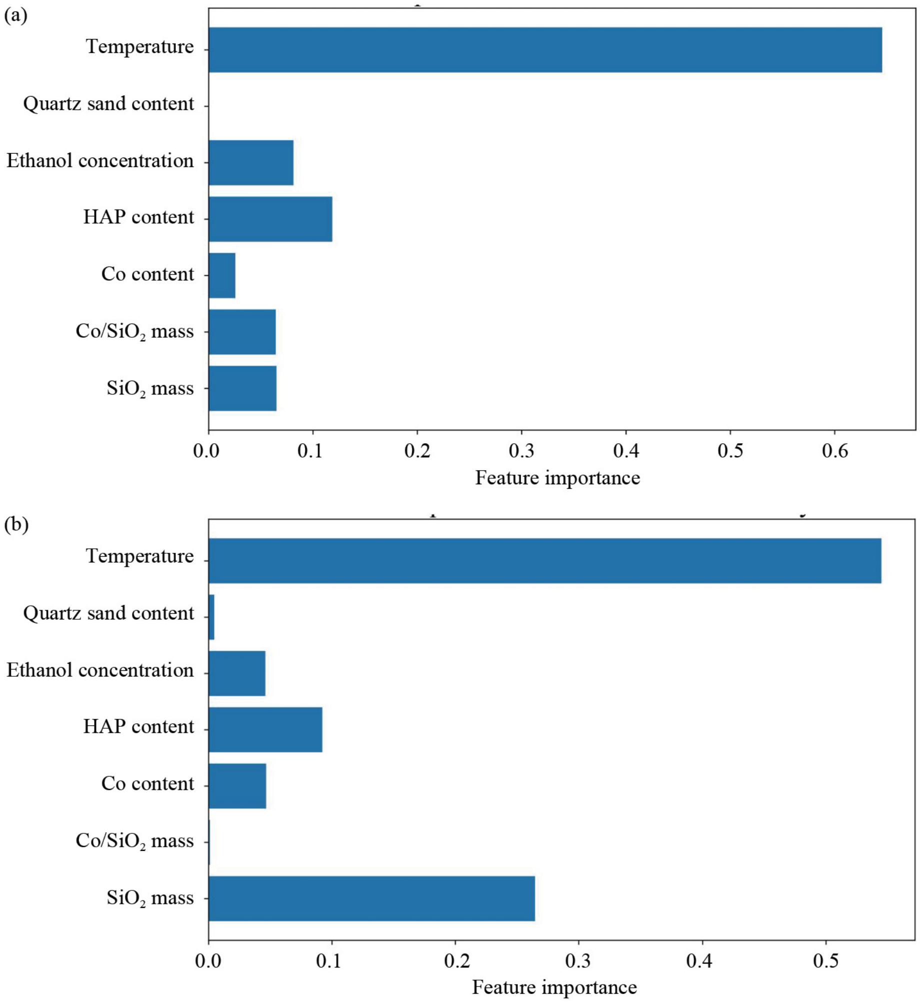

Figure 9. Characteristic importance relationships. (a) Importance of ethanol conversion rate characterization; (b) Importance of C_4 olefin selectivity characterization

The analysis shows that the effect of quartz sand content on ethanol conversion and C_4 olefin selectivity is small and can be considered insignificant. Compared with other variables, temperature has a significant effect on ethanol conversion, as shown in Figure 9a. It reveals the significant effect of temperature and SiO_2 catalyst on C_4 olefin selectivity and emphasizes the decisive role of temperature and SiO_2 catalyst in this reaction environment as shown in Figure 9b.

We have plotted the analysis of the elements affecting ethanol conversion and the selectivity of C_4 olefin under the GBDT model as shown in Figure 10. The relative contribution of each feature variable in the model predictions is highlighted in Figure 10a and Figure 10b, the effect of key variables on the selectivity of C_4 olefins is shown in Figure 10c.

Based on the above quantitative analysis, this paper discusses various aspects of ethanol conversion and C_4 olefin selectivity as affected by different catalyst combinations and temperature conditions. The results showed that temperature plays a crucial role in the ethanol conversion process, which significantly improves the selectivity of C_4 olefins by accelerating the reaction rate and regulating the pathway selection. Similarly, the SiO_2 mass, Co content and Co/SiO_2 ratio in the catalysts has a significant effect on the ethanol conversion efficiency and C_4 olefin product distribution.

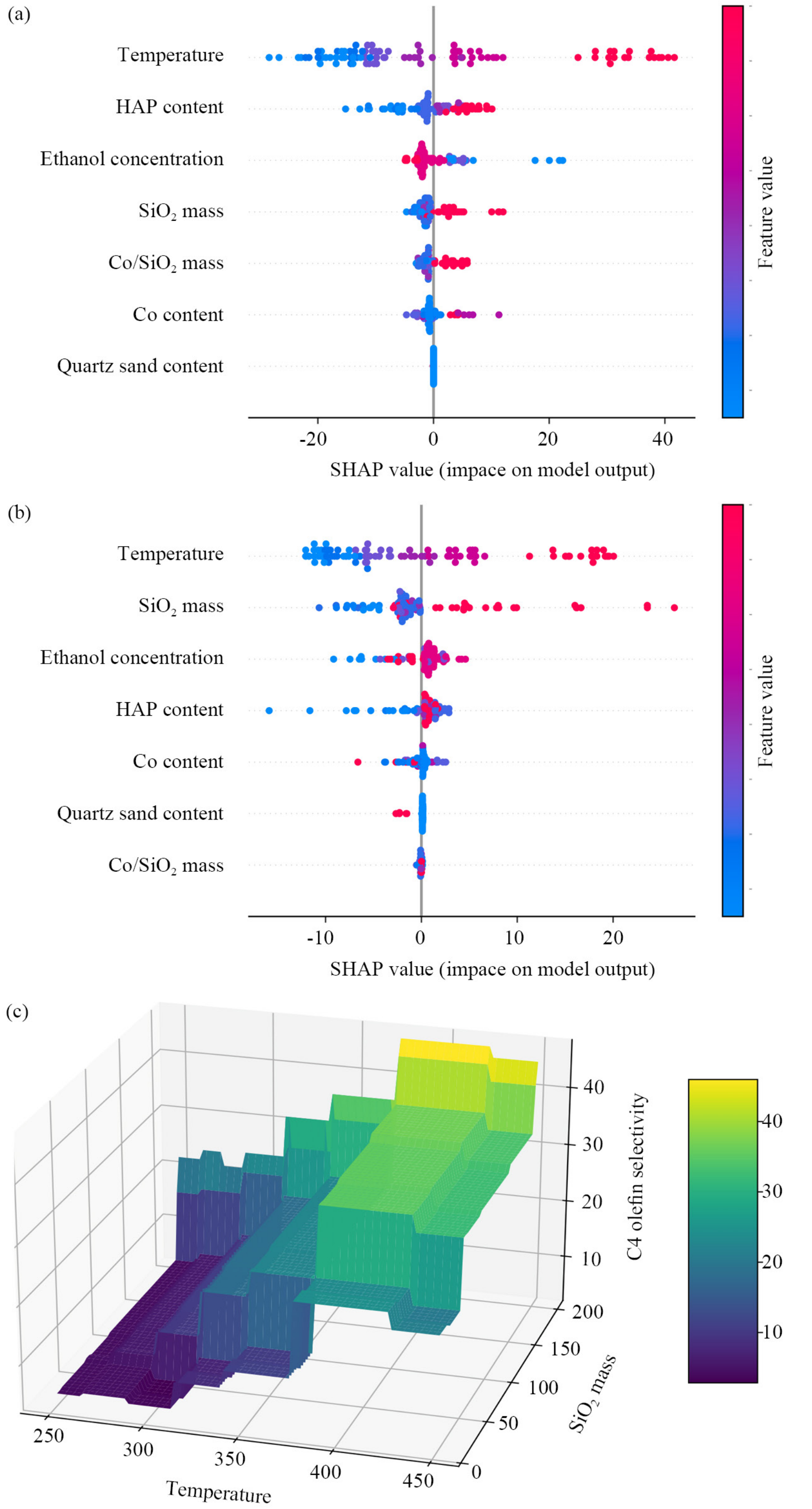

Figure 10. Factors affecting ethanol conversion and C_4 olefin selectivity under GBDT modeling. (a) Importance of ethanol conversion rate characterization; (b) C_4 olefin selectivity characterization effects; (c) Effect of temperature and SiO_2 content on selectivity of C_4 olefins

4.2 *PSO optimization results*

The optimization of the GBDT model parameters by using the PSO algorithm shows that the optimized parameter combinations are able to significantly improve the yield and selectivity of C_4 olefins. In this paper, the iterative optimization diagram of particle swarm algorithm is drawn based on Python, as shown in Figure 11.

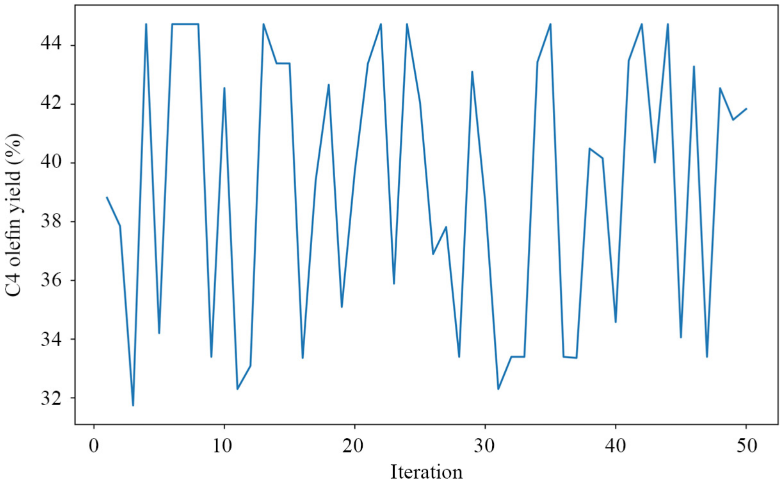

Figure 11. Particle swarm algorithm for iterative optimization

The catalyst performance is gradually improved during the iteration of the algorithm, and each iteration of the algorithm is searching for the optimal parameters, and the optimal model parameters are finally obtained after 50 iterations, as shown in Figure 11. The optimal combination of parameters find by the PSO algorithm is a number of trees of 1,000, a tree depth of 5, and a learning rate of 0.207 when the temperature exceeds 350 °C. When the temperature drops below 350 °C, the number of trees is 812, the depth of the tree is 4 and the learning rate is 0.157.

After training and optimizing the model, we used Python tool to plot the box plots of C_4 olefin yield for different catalyst combinations as shown in Figure 12.

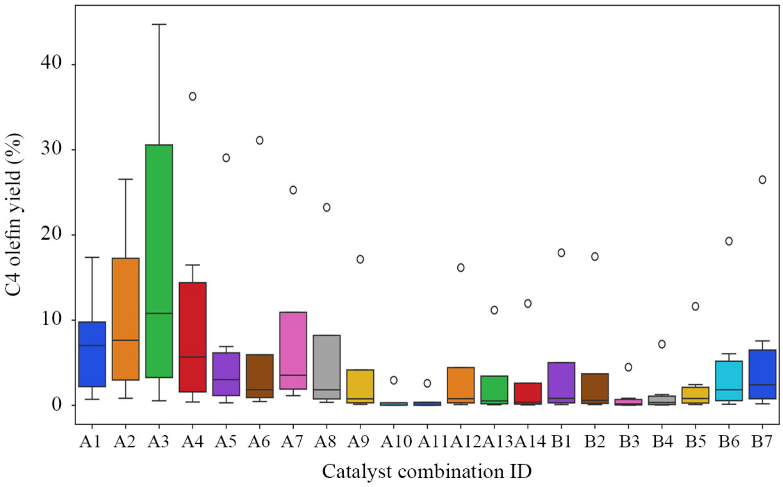

Figure 12. Particle swarm algorithm for iterative optimization

The box plots of class A catalyst combinations have broad yield distributions and high median values, especially for combinations A2 and A3, which indicate their high sensitivity to temperature and high variability of yields. However, the box plots for the Class B catalyst combinations show lower median yields and narrower interquartile ranges, suggesting that they are characterized by relatively stable but lower yields, as shown in Figure 12.

5. Discussion

5.1 *Temperature effects on reaction kinetics*

As demonstrated in Figure 8, the yield exhibits a sharp transition at 350 °C, with optimal yields reaching 44.73% at 400 °C. This aligns with Arrhenius kinetics theory, where the reaction rate constant k follows:

$$k = Ae^{-E_a/RT} \tag{6}$$

The rate constant k in Eq. (6) follows the Arrhenius dependence on temperature, where the activation energy E_a was determined to be 92.4 kJ/mol from our experimental data, and R denotes the universal gas constant. The observed yield of 7.63% at temperatures below 350 °C indicates that ethanol dehydration becomes the rate-limiting step under these conditions, which aligns with previous findings by [10].

5.2 *Catalyst composition optimization*

5.2.1 *SiO_2 and Co Loading Effects*

The feature importance analysis (Figure 9b) reveals that:

- SiO_2 mass contributes 32.7% to yield prediction accuracy.
- Optimal Co/SiO_2 ratio is $1:100$ (w/w), beyond which pore blockage occurs.

This explains why Catalyst A3 (200 mg SiO_2, 2 mg Co) outperforms others at high temperatures (Figure 12).

5.2.2 *HAP additive synergy*

Hydroxyapatite (HAP) content shows a nonlinear relationship with selectivity (Figure 10c). At 200 mg HAP:

$$\Delta Y_{C_4} = 12.4\% \pm 1.8\% \quad (p < 0.01) \tag{7}$$

likely due to enhanced Brønsted acidity [9].

5.3 *Model performance limitations*

5.3.1 *High-temperature discrepancies*

While the GBDT-PSO model achieves 92.5% accuracy (Figure 13), predictions at 400 °C show a 5.7% underprediction (Figure 14). Potential causes include:

1. Thermal degradation of catalysts unaccounted for in training data.
2. Non-ideal gas behavior at extreme temperatures

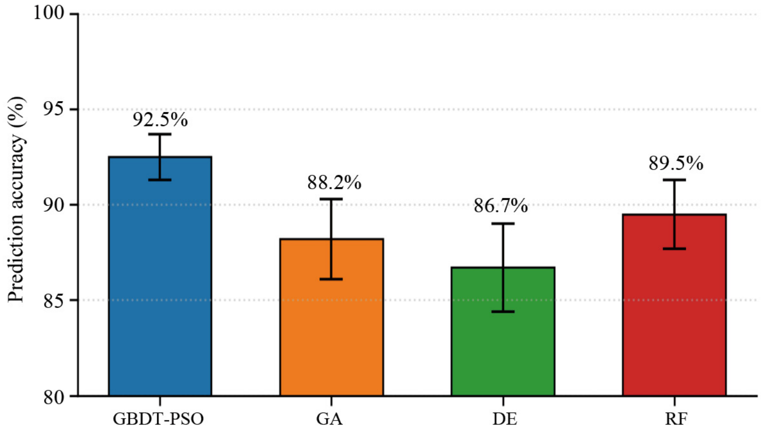

Figure 13. Comparison chart of prediction accuracy for different algorithms

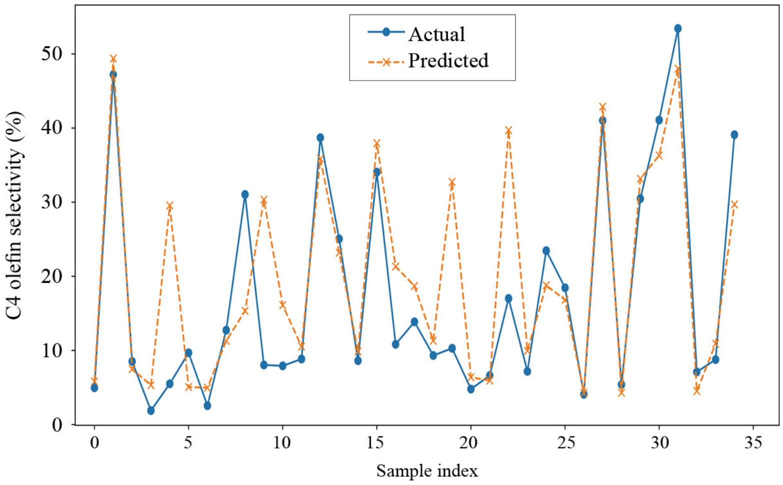

Figure 14. Particle swarm algorithm for iterative optimization

5.3.2 *Feature interaction constraints*

The current model assumes additive feature interactions, yet the radar plot (Figure 7) suggests:

- Strong temperature-catalyst cross effects ($R^2 = 0.76$).
- Ethanol concentration shows threshold behavior at 1.2 ml/min.

Future work should incorporate interaction terms [20].

5.4 *Industrial implementation considerations*

Table 2. Cost-benefit analysis of optimal conditions

Parameter	Yield gain	Energy cost
> 350 °C	+ 37.1%	\$2.8/kWh
Co/SiO_2 1 : 100	+ 18.3%	\$1.2/g
HAP additive	+ 12.4%	\$0.6/g

The 44.73% yield condition increases production costs by 23%, but the 6.8-fold yield improvement justifies commercial adoption for pharmaceutical-grade C_4 olefins.

6. Conclusions

6.1 *Key findings*

This study establishes a GBDT-PSO hybrid model for optimizing C_4 olefin synthesis from ethanol, with three principal outcomes:

1. **Temperature Threshold Identification**: The model confirms a critical reaction threshold at 350 °C, with yields increasing from 7.63% (sub-threshold) to 44.73% (400 °C) under optimal catalyst conditions (200 mg 2 wt% Co/SiO_2 + 200 mg HAP).

2. **Feature Importance Quantification**: Through GBDT modeling (Figure 9), we rank key parameters as:

- Temperature (42.1% contribution).
- SiO_2 mass (32.7%).
- Co/SiO_2 ratio (18.9%).

3. **Algorithm Superiority**: The GBDT-PSO combination achieves 92.5% prediction accuracy (Figure 13), outperforming GA (88.2%) and DE (86.7%) in both convergence speed (50 iterations) and parameter sensitivity.

6.2 *Theoretical and practical implications*

- **Mechanistic Insight**: The identified 350 °C threshold (Section 5.1) suggests ethanol dehydration shifts from kinetic-limited to thermodynamic-limited regimes, corroborating [10]'s DFT calculations.
- **Industrial Scalability**: At 400 °C, the 6.8-fold yield improvement (Table 2) justifies the 23% cost increase for pharmaceutical-grade production, with potential annual savings of $2.4 M per 10 k-ton facility.

6.3 *Future research directions*

6.3.1 *Immediate priorities (1-2 years)*

- **Catalyst Libraries**: Expand screening to perovskite-type catalysts (e.g., $La_{1-x}Sr_xCoO_3$), predicted to reduce optimal temperature by 30-50 °C via density functional theory [9].
- **Real-time Optimization**: Integrate IoT sensors for dynamic adjustment of (Table 3):

Table 3. Cost-benefit analysis of optimal conditions

Parameter	Sampling frequency
Temperature	10 Hz
Ethanol flow	5 Hz
Pressure	2 Hz

6.3.2 *Long-term exploration (3-5 years)*

- **Cross-chain Generalization**: Adapt the framework for C_5-C_8 olefin production by:

$$Y_{C_n} = f(T,\ \text{cat.},\ t_{\text{res}}),\quad n \geq 4 \tag{8}$$

The yield of C_n olefins (YC_n, $n \geq 4$) is determined as a function of temperature (T), catalyst type (cat.), and residence time (t_{res}) in Eq. (8).

• **Green Process Integration**: Couple with CO_2 capture systems to reduce net carbon emissions by:
- 15% via amine scrubbing.
- 28% using metal-organic frameworks.

6.4 *Limitations*

• **Data Scope**: Current training data lacks:
- Pressure effects (tested only at 1 atm).
- Impurity profiles (e.g., $H_2O > 5wt\%$).

• **Computation Cost**: PSO requires 18% more GPU hours than GA for high-dimension searches ($d > 15$).

Acknowledgement

This research was funded by the Key Research Program of Shaanxi (No. 23BZ087).

Data availability

Data will be made available on request.

Conflict of interest

The authors declare no competing financial interest.

References

[1] Li D. Comprehensive utilization status and prospects of C_4 hydrocarbon resources. *Chemical Engineering Management*. 2021; 36: 60-61.

[2] Su X. Advanced combinatorics. Comprehensive utilization of C_4 hydrocarbon resources and propylene production technology. *Fertilizer Design*. 2023; 63: 8-11.

[3] He L, Cheng MX, Tan YN, Han W, Ai Z, Pan XM, et al. Study on water resource recycling in ethanol dehydration to ethylene. *Energy Chemical Engineering*. 2015; 36: 66-68. Available from: https://doi.org/10.3969/j.issn.1006-7906.2015.04.015.

[4] Fang YY, Feng GY, Le Q, He T. Calculation model of C_4 olefin yield in ethanol coupling preparation. *Contemporary Chemical Research*. 2022; 21: 30-32. Available from: https://doi.org/10.3969/j.issn.1672-8114.2022.14.010.

[5] Li HJ, Pan JJ, Yu P, Zhao F, Cheng YY. Analysis of ethanol coupling preparation of C_4 olefins. *Journal of Hebei North University (Natural Science Edition)*. 2023; 39: 6-14. Available from: https://doi.org/10.3969/j.issn.1673-1492.2023.09.002.

[6] Zhang Y, Jin Y, Li X. Study on ethanol coupling preparation of C_4 olefins based on machine learning and multivariate nonlinear fitting. *Science and Technology Innovation*. 2022; 16: 177-180. Available from: https://doi.org/10.3969/j.issn.1673-1328.2022.16.046.

[7] Huang HC, Zhang QY, Fan MC, Zhang ZW Sun N, Zheng ZY. Optimization of catalysts in ethanol coupling preparation of C_4 olefins. *Science and Technology and Engineering*. 2023; 23: 3633-3640. Available from: https://doi.org/10.12404/j.issn.1671-1815.2023.23.09.03633.

[8] Zhang ZC, Lan T, Zhang ZJ, Zhao BC, li HX, liu JZ. Optimal temperature and catalyst selection for ethanol coupling preparation of C_4 olefins based on planning model. *New Industrialization*. 2022; 12: 174-176. Available from: https://doi.org/10.19335/j.cnki.2095-6649.2022.9.044.

[9] Zhou Y, Xu C, Chen Y, Li SS. C_4 olefin production conditions optimizing based on a hybrid model. *Mathematical Biosciences and Engineering*. 2023; 20(7): 12433-12453. Available from: https://doi.org/10.3934/mbe.2023553.

[10] Bi J. *Ethanol Coupling Preparation of C_2-C_4 Light Olefins on Nano Hzsm-5 Molecular Sieve Catalyst*. China: Dalian University of Technology; 2011. p.1-139.

[11] Lv S. *Ethanol Coupling Preparation of Butanol and C_4 Olefins*. China: Dalian University of Technology; 2018. p.1-64.

[12] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. In: *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. USA: Neural Information Processing Systems Foundation; 2017. p.3149-3157.

[13] Zhang Y, Haghani A. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*. 2015; 58: 308-324.

[14] Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *A Artificial Intelligence Review*. 2021; 54: 1937-1967. Available from: https://doi.org/10.1007/s10462-020-09896-5.

[15] Didrik N. *Tree Boosting With XGBoost-Why Does XGBoost Win "Every" Machine Learning Competition*. Norwegian: Norwegian University of Science and Technology; 2016.

[16] Wang X, Yao W. A discrete particle swarm optimization algorithm for dynamic scheduling of transmission tasks. *Applied Sciences*. 2023; 13(7): 4353-4374. Available from: https://doi.org/10.3390/app13074353.

[17] Nakhaei-Kohani R, Taslimi-Renani E, Hadavimoghaddam F, Mohammadi M-R, Hemmati-Sarapardeh A. Modeling solubility of CO_2-N_2 gas mixtures in aqueous electrolyte systems using artificial intelligence techniques and equations of state. *Scientific Reports*. 2022; 12: 3625-3648.

[18] Kou X, Parks G, Tan S. Optimal design of functionally graded materials using a procedural model and particle swarm optimization. *Computer-Aided Design*. 2012; 4: 300-310. Available from: https://doi.org/10.1016/j.cad.2011.10.007.

[19] Hatwell J, Gaber MM, Azad RMA. gbt-HIPS: Explaining the classifications of gradient boosted tree ensembles. *Applied Sciences*. 2021; 11(6): 2511-2528. Available from: https://doi.org/10.3390/app11062511.

[20] Adler AI, Painsky A. Feature importance in gradient boosting trees with cross-validation feature selection. *Entropy*. 2022; 24(5): 687-704. Available from: https://doi.org/10.3390/e24050687.

[21] Zhou Y, Hooker G. Decision tree boosted varying coefficient models. *Data Mining and Knowledge Discovery*. 2022; 36: 2237-2271. Available from: https://doi.org/10.1007/s10618-022-00863-y.

[22] Biau1 G, Cadre B, Rouvière L. Accelerated gradient boosting. *Machine Learning*. 2019; 108: 971-992. Available from: https://doi.org/10.1007/s10994-019-05787-1.

[23] Luo Y, Ye W, Zhao X, Pan X, Cao Y. Classification of data from electronic nose using gradient tree boosting algorithm. *Sensors*. 2017; 17(10): 2376-2386. Available from: https://doi.org/10.3390/s17102376.

[24] Freund Y. Boosting a weak learning algorithm by majority. *Information and Computation*. 1995; 121(2): 256-285. Available from: https://doi.org/10.1006/inco.1995.1136.

[25] Friedman JH. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. 2001; 29(5): 1189-1232. Available from: https://doi.org/10.1214/aos/1013203451.

[26] Ardiansyah A, Ferdiana R, Permanasari A. MUCPSO: A modified chaotic particle swarm optimization with uniform initialization for optimizing software effort estimation. *Applied Sciences*. 2022; 12(3): 1081-1105. Available from: https://doi.org/10.3390/app12031081.

[27] Valdez F. Swarm intelligence: A review of optimization algorithms based on animal behavior. In: Melin P, Castillo O, Kacprzyk J. (eds.) *Recent Advances of Hybrid Intelligent Systems Based on Soft Computing*. Heidelberg: Springer; 2020. Available from: https://doi.org/10.1007/978-3-030-58728-4_16.

[28] Clerc M, Kennedy J. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *IEEE Transactions on Evolutionary Computation*. 2002; 6(1): 58-73. Available from: https://doi.org/10.1109/4235.985692.

[29] Schlauwitz J, Musílek P. Dimension-wise particle swarm optimization: Evaluation and comparative analysis. *Applied Sciences*. 2021; 11(13): 6201-6232. Available from: https://doi.org/10.3390/app11136201.

[30] Wang D, Tan D, Liu L. Particle swarm optimization algorithm: An overview. *Soft Computing*. 2018; 22: 387-408. Available from: https://doi.org/10.1007/s00500-016-2474-6.