MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **91** (2024) 557–592
ISSN: 0340–6253
doi: 10.46793/match.91-3.557G

OCS-TGBM: Intelligent Analysis of Organic Chemical Synthesis Based on Topological Data Analysis and LightGBM

Yanhui Guo[1,*], Lichao Peng[2,*], Zixin Li[1], Meng'en Qin[1], Xue Jiao[3], Yun Chai[4,†], Xiaohui Yang[1,†]

[1] *Henan Engineering Research Center for Artificial Intelligence Theory and Algorithms, School of Mathematics and Statistics, Henan University, Kaifeng, China, 475000*

[2] *National and Local Joint Engineering Research Center for Applied Technology of Hybrid Nanomaterials, Henan University, Kaifeng, 475000*

[3] *Henan Institute of Science and Technology, Eastern HuaLan Avenue, Xinxiang City, Henan, China, 453003*

[4] *College of Chemistry and Chemical Engineering, Henan University, Kaifeng, China, 475000*

chaiyun@henu.edu.cn, xhyanghenu@163.com

(Received June 11, 2023)

Abstract

Organic synthesis has been widely used in drug discovery and development. The intelligent prediction and analysis of high-throughput coupling reaction yield is one of the important and challenging research hotspots in the field of organic synthesis. However, the existing methods focus on intelligent prediction rather than study and interpret the internal relationship between reaction conditions and yield. For tackling this problem, an intelligent analysis organic chemical synthesis model by combining topological data analysis (TDA) and Light Gradient Boosting Machine (LightGBM), named

[*] Yanhui Guo and Lichao Peng contributed equally to this work.
[†] Corresponding author.

OCS-TGBM, is proposed to deeply explore the internal relationship between reaction conditions and yield, and obtain high-yield reaction conditions and combinations. In order to further enhance the performance of the OCS-TGBM model, a stratified diversity sampling strategy is introduced. Experimental results show that the OCS-TGBM model is superior to other methods in analyzing and predicting the reaction performance of high-throughput organic chemical synthesis. And it provides intelligent assistance for the optimal design of the reaction system and the evaluation of reaction conditions, thus greatly accelerating the process of the drug discovery and development.

1 Introduction

Organic synthesis plays a vital role in the innovative research and development of new drugs as well as the optimization of old drugs. With the development of organic synthesis technology, the coupling reaction catalyzed by transition metals is also developing rapidly. Among them, the cross-coupling reaction catalyzed by palladium (Pd) is a widely used method with high efficiency, excellent selectivity, and mild reaction conditions, which is an effective tool for modern organic synthesis.

Buchwald-Hartwig amination reaction is one of the research hot points in the field of coupling reaction of C-N bond catalyzed by Pd [1–4]. In order to improve the yield of Buchwald-Hartwig cross coupling reaction, researchers have been committed to improving reaction conditions such as ligands and additives in the reaction [5–8]. However, the current Buchwald-Hartwig cross coupling reaction is obviously facing corresponding shortcomings. For example, the reaction conditions are harsh, the synthetic route is complex, the reaction time is long, the solvent pollutes environment, high cost and difficult to achieve [9]. Therefore, designing green, simple, and efficient chemical synthesis method has become the focus of research on Buchwald-Hartwig cross coupling reaction.

In recent years, machine learning (ML) as an efficient method has been gradually applied in the field of bioinformatics and chemistry [10, 11]. It shows more and more competitiveness in the research of chemical reaction prediction [12–14], drug performance prediction [15–19], screening for target compounds [20–23], molecular material design [24–26]. Recently, researchers considered using ensemble tree models to predict the performance of chemical reactions, which are easy to analyze and interpret. For example, the Random Forest model is used to predict the toxicity of chemicals [27,28], the stereoselectivity of glycosylation [29]. And hybrid genetic algorithm decision tree model is used to predict the effect of solvent structure on the reaction rate [30].

In 2018, Ahneman et al. [12] reported the prediction of Buchwald-Hartwig amination reaction yield by Random Forest model, which is an advanced study of ML method in the field of multidimensional chemical spatial prediction. The yield was predicted with an accuracy of $R^2 = 0.92$, $RMSE = 7.80$. However, the data obtained by Ahneman et al. [12] is high-dimensional data, and it has lots of redundant information, the Random Forest algorithm cannot make prediction beyond the training data range when solving the regression problem, which will lead to over-fitting.

Based on the above, our team proposed an integrated feature selection based on importance and relevance, then obtained comprehensive and concise feature descriptors data (21 feature descriptors) [36,37]. And XGBoost was introduced to improve the precision. However, XGBoost is relatively large time cost due to the pre-sorted algorithm.

Furthermore, researchers not only want good predictions, but also aim to delve deeper into the relationship between reaction conditions and yield. Cluster analysis, as a primary method of data mining, has received widespread attention. In 2000, H. Edelsbrunner et al. proposed the Topological Data Analysis (TDA) model [31]. It's sensitive to both large-scale and small-scale patterns, which other analysis methods such as PCA and K-Means may not be able to detect [32, 33]. It also can find some small categories that cannot be found by traditional methods. Therefore, this method has played a great role in the field of gene and cancer research [34, 35].

Motivated by these works, this paper focuses on exploring the relationship between reaction conditions and yield, and improving the intelligent prediction yield model using the feature descriptors screened by our team. The main contributions are as follows,

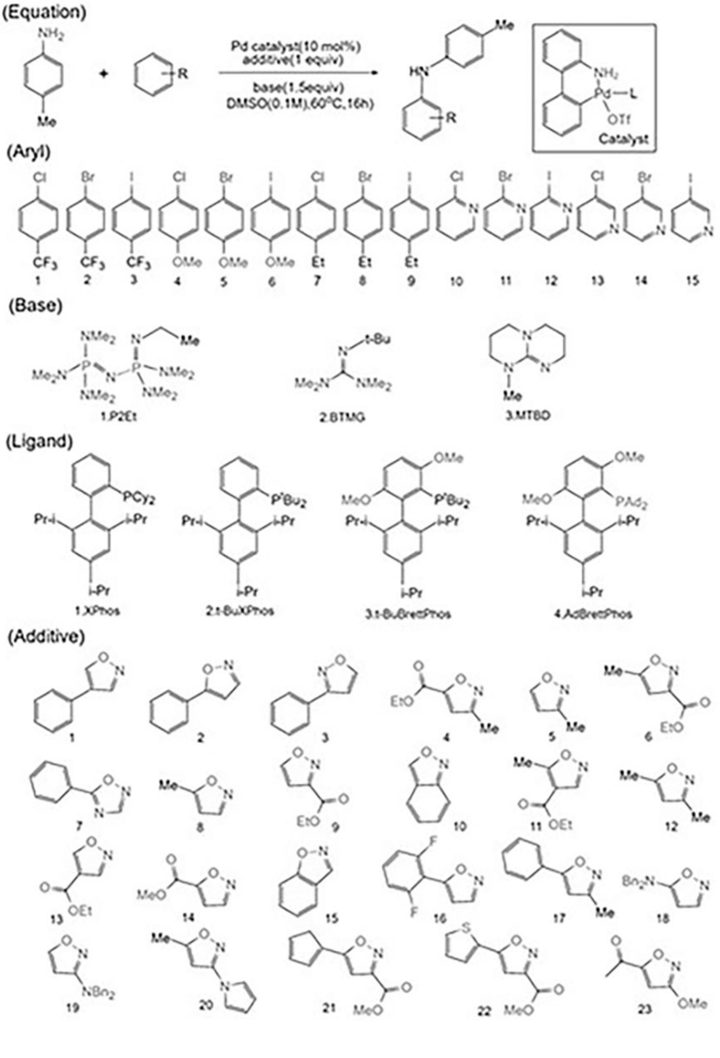


Figure 1. All reaction components of Buchwald-Hartwig amination reaction.

(1) We propose OCS-TGBM, an intelligent analysis organic chemical synthesis model by combining topological data analysis (TDA) and Light Gradient Boosting Machine (LightGBM). OCS-TGBM can be used to deeply explore the intrinsic relationship between reaction conditions and yield, and make intelligent predictions. Besides, it can reduce running

time while improve the accuracy of intelligent prediction.

(2) The stratified diversity sampling is proposed to divide the training set and testing set, it enhances the performance of the model.

(3) Experimental results show that the OCS-TGBM model is superior to other methods in analyzing and predicting the reaction performance of high-throughput organic chemical synthesis. That is, the OCS-TGBM is an effectiveness model.

2 Intelligent analysis and prediction model for reaction yield—OSC-TGBM

To investigate the relationship between reaction conditions and yield in depth, and to achieve efficient intelligent prediction, this paper proposes an intelligent organic chemistry synthesis system called OCS-TGBM. OCS-TGBM combines Topological Data Analysis (TDA) and Light Gradient Boosting Machine (LightGBM). First, TDA is applied to comprehensively analyze the relationship between reaction conditions and yield. Then, LightGBM is utilized to achieve efficient intelligent prediction. The introduction of OCS-TGBM provides an innovative approach for research and synthesis processes in the field of organic chemistry, with the hope of further enhancing reaction efficiency and product selectivity.

Nextly, this section will introduce TDA, LightGBM, stratified diversity sampling strategy and three evaluating indicators.

2.1 TDA-based hidden information mining model for high-dimensional data

The unique function of TDA make it have broad potential in the field of data analysis and mining, which can widely explore and understand the complex high-dimensional data spaces. The main methods of TDA include persistent homology and Mapper. Mapper helps data analysts summarize and visualize complex datasets, providing intuitive insights into the data [38].

2.1.1 Mapper algorithm

The Mapper summarizes the topological structure of the datasets into a graph through a mapping $f : X \to G$. It is a way of constructing graphs from data, which reveals the topological characteristics of high-dimensional data space.

The Mapper algorithm is divided into sequential Mapper and distributed Mapper, among which distributed Mapper is widely used. In order to ensure that the output of the distributed Mapper is the same as that of the sequential Mapper, some coverage preprocessing is required to obtain the final Mapper output.

For coverage preprocessing, first construct an N-chain coverage of $[a, b]$, $[a, b]$ is covered by N open intervals $A_1, A_2 \cdots A_N$. When $|i - j| = 1$ and $|i - j| = \phi$, $A_{i,j} := A_i \cap A_j \neq \phi$. Then construct an open cover U_i for each open set A_i, $\{U_i\}_{i=1}^{N}$ coverage meets the following conditions,

(1) $A_{i,i+1}$ is an open set covering of U_i and U_{i+1}, that is $U_i \cap U_{i+1} = \{A_{i,i+1}\}$.

(2) If $U_i \in u_i$ and $U_{i+1} \in u_{i+1}$ make $U_i \cap U_{i+1} \neq \phi, i = 1, 2, \cdots N - 1$, there is $U_i \cap U_{i+1} = {A_{i,i+1}}_i$.

For the set of $\{A_i, U_i\}_{i=1}^{N}$, where $\{A_i\}_{i=1}^{N}$ is the N-chain coverage of $[a, b]$ and U_i is the coverage of A_i.

Sequential Mapper algorithm. Given a finite cover $u = \{U_1, U_2, \ldots U_k\}$ of $f(X)$, the cluster $X_{i,j} \subset X_i$ for each set $X_i := f^{-1}(U_i)$ is computed by using clustering algorithm.

Distributed Mapper algorithm. After preprocessing the coverage and obtaining the set $\{A_i, U_i\}_{i-1}^{N}$, firstly, map each pair of (A_i, U_i) to a specific processor P_i. Then determine the set $X_i \subset X$ of points, it is mapped to A_i by f and simultaneously run the sequential Mapper construction on cover $(f|x_i) * (u_i), (i = 1, 2, \cdots N)$, thus, obtaining N graphs G_1, G_2, G_N, if $N = 1$, return graph G_1. Then, let $C_{j1}^i, C_{j2}^i, \cdots C_{ji}^i$ be the cluster obtained from $f^{-1}(A_{i,i+1})$. By selecting coverings u_i and u_{i+1}, these clusters are represented by vertices $v_{j1}^i, v_{j2}^i, \cdots, v_{ji}^i$ in G_i and G_{i+1}(each v_k^i corresponds to cluster C_k^i). Finally, by constructing $A_{i,i+1}$, u_i and u_{i+1}, each $f^*(u_i)$ and $f^*(u_{i+1})$ share a cluster C_{jk}^i in $f^*(A_{i,i+1})$. So C_{jk}^i is represented by a vector in graphs G_i and G_{i+1}, and by considering disjoint joint graphs $G_1 \cup G_2 \cup \cdots \cup G_N$ when merging, then take the quotient of this graph to determine the corresponding vertices in G_i and G_{i+1} $(1 \leq i \leq N-1)$. Thus, the subgraphs $G_1, G_2, \cdots G_N$ are merged into a graph G.

2.1.2 Clustering for topological data analysis

The main steps of TDA clustering visualization are as follows,

(1) Calculating a filtered value for each data point by using a filter function. The $L - \inf inity$ is used as filter function in this paper. (The value of $L - \inf inity$ is the distance from the point to the point farthest from it, which is a centrality indicator.)

$$L - \inf inity = \max_{j=1,\cdots len(d)} \sqrt{\sum_{k=d[j][0]}^{d[j][20]} \sum_{l=d[i][0]}^{d[i][20]} (k-l)^2}, \tag{1}$$

where d is the original data, $len(d)$ is sample size, and n is the number of features, $d[j]$ is the jth sample, $d[j][0]$ is the first feature of the jth sample.

(2) The data points are divided into different filter value intervals from small to large according to their filter values. However, it should be noted that adjacent filter value intervals are set with a certain overlapping area, that is, the points in the overlapping area belong to two intervals at the same time. The set of N equal-length intervals are determined by the two resolution parameters (N intervals and p overlap percentages).

(3) Clustering the data in each interval. In this paper, single-link cluster is used to cluster each group. Let N is the number of points in the box. This paper first constructs a single-link dendrogram for the data in bins and records the threshold for each transition in the cluster. Then selecting an integer K, and constructing a K-interval histogram for these transition values. The last threshold before the first gap in the histogram is used for clustering. Note that the larger values of K, the more clusters are produced, and the lower values of K, the fewer clusters are produced.

(4) Putting together the subclasses obtained by each interval clustering in the previous step, and each subclass is represented by a circle of different size. If two categories have common original data points (this is why the

intervals need to overlap each other), add an edge between them.

2.2 Construction of a yield prediction model based on LightGBM

LightGBM is an open-source framework based on the Gradient Boosting Decision Tree (GBDT). LightGBM is proposed to improve the performance of GBDT in processing massive data. The histogram-based decision tree algorithm and distributed processing make GBDT better and faster to apply to big data analysis, such as industry, medical treatment and so on. Therefore, this section will introduce GBDT and the improved LightGBM model based on GBDT [39].

The algorithm goal of GBDT is to optimize the loss function $L(\varphi) = \sum l(\hat{y}_i, y_i)$. The idea is too iterative generate multiple weak models, and then adding up the prediction results of each weak model. The latter model $f_t(x)$ is generated based on the effect of the previous learning model $f_{t-1}(x)$. Assuming the GBDT model contains K weak learners, and let A and B are the parameters of the classifier, there is,

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i; \alpha_t). \tag{2}$$

Therefore, the objective function $L(\varphi) = \sum l(\hat{y}_i, y_i)$ of GBDT can be transformed into the following form,

$$L^{(t)} = \sum_{i=1}^{n} l\left[y_i, \hat{y}_i^{(t-1)} + f_t(x_i; \alpha_t)\right]. \tag{3}$$

Next, through the first-order Taylor expansion, removing the constant term, and optimizing the loss function term to optimize the GBDT objective function. Let the first-order derivative be $g_i = l'(y_i, \hat{y}_i^{(t-1)})$, then the first-order Taylor expansion is,

$$l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) \approx l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i; \alpha_t). \tag{4}$$

Substitute further into the objective function (1) to get $L^{(t)}$,

$$L^{(t)} = \sum_{i=1}^{n} \Big[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i; \alpha_t) \Big], \tag{5}$$

it can be found that $l(y_i, \hat{y}_i^{(t-1)})$ is the loss corresponding to the previous step $t-1$. If A is set, it can be guaranteed that the subtraction of the last part of the formula must be a positive number. So, taking a negative gradient in this way will make the loss decrease step by step. The parameters that minimize the objective function are,

$$\beta_t, \alpha_t = \arg\min \sum_{i=1}^{n} \Big[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i; \alpha_t) \Big], \tag{6}$$

where β_t, α_t are the parameters of the direction that makes the loss function of the model $f_{t-1}(x)$ of the previous step decreases the fastest, and they are also the parameters to get the direction of model $f_t(x)$. Because the direction where the loss function of $f_{t-1}(x)$ decreases the fastest is $-g_i = -l'(y_i, \hat{y}_i^{(t-1)})$, α_t, β_t are obtained by the least squares method,

$$\begin{aligned} \alpha_t &= \arg\min \sum_{i=1}^{n} \left[-g_i - g_i f_t(x_i; \alpha) \right]^2 \\ \beta_t &= \arg\min \sum_{i=1}^{n} \Big[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i; \alpha_t) \Big]. \end{aligned} \tag{7}$$

Thus, the final model can be obtained,

$$f_t(x) = f_{t-1}(x) + \beta_t f_t(x; \alpha_t). \tag{8}$$

Gradient Boosting tree is an optimization process that uses the additive model and forward distribution algorithm for learning. When building decision trees in GBDT, it is fitted with a negative gradient, and when calculating the information gain, all simples need to be scanned to find the optimal splitting point, which greatly reduces the efficiency of the model. LightGBM solves this problem through the Gradient-based One-side Sampling (GOSS) algorithm and the Exclusive feature bundling (EFB) algorithm, they are the core algorithm of LightGBM.

The GOSS algorithm attempts to solve this problem from the perspective of reducing the sample size. By sorting the absolute values of the data gradients, retaining instance with larger gradients among a and set it as data subset A. And randomly sampling b instances from the remaining small gradient instances as the data subset B. The information gain is calculated based on the samples collected from the gradients, which greatly reducing the amount of calculation and ensures accuracy. The specific process of GOSS are as follows,

(1) Firstly, making prediction according to the model, and the sample prediction value $preds$ is obtained.

(2) Calculating $loss$ according to $preds$, then further calculating the sample gradient, and the initial assignment of sample weight w is equal to 1.

(3) According to the absolute value of the sample gradient, the sequence $sorted$ is obtained by descending sort, which is the index array of the samples.

(4) Large gradient sample data, selecting $topN = a * len(I)$ to get $topSet$, which is also an index array.

(5) Small gradient sample data, randomly selecting $randN = b * len(I)$ from the remaining samples to get $randSet$.

(6) Combining rows $topSet$ and $randSet$ to get $usedSet$, the size is equal to $(a + b) * len(I)$.

(7) Multiply the sample weight of the small sample by the weight coefficient factor $(1 - a)/b$ to get the new sample weight w.

(8) According to the sample I on index $usedSet$, the gradient g, and the weight w, a new weak learner $newModel$ is obtained.

(9) Adding the new weak learner $newModel$ to the total model (LightGBM is an additive model).

Where I is the training data, a is the sampling ratio of large gradient data, and b is the sampling ratio of small gradient data.

The EFB algorithm achieves dimensionality reduction and improves efficiency by bundling mutually exclusive features. For example, for one-hot encoded features, the features cannot have non-zero values at the same time. These mutually exclusive features are bundled, and the bundled features are merged to construct a feature histogram that is equivalent to a single feature, thus reducing computational costs. The greedy strategy for feature bundling in EFB can be summarized as follows,

(1) The features are taken as the vertices of the graph, and the non-mutually exclusive features are connected (there exist samples that are not zero at the same time). And the number of samples with features that are not zero at the same time is taken as the weight of the edge.

(2) Sorting the features in descending order according to the degree of the vertices. The greater the degree, the greater the conflict between the feature and other features (less likely to be bundled with other features).

(3) Setting a maximum conflict threshold K, the outer loop first iterates over each of the sorted features mentioned above, and then iterates over the existing feature bundles/clusters. If it is found that adding the feature to a particular cluster would not exceed the maximum threshold K of conflicts, then add the feature to that cluster. Otherwise, creating a new feature cluster and add the feature to the newly created cluster.

Finally, since histogram-based algorithms store discrete bins rather than continuous feature values, feature bundles can be constructed by adding exclusive features that reside in different containers, which is achieved by adding an offset to the original values of the features. This is the Merge Exclusive Features of LightGBM.

2.3 Stratified diversity sampling

Stratified sampling can improve the representativeness of samples, and diversity sampling can make the model better for feature learning.

The main steps of stratified diversity sampling are as follows,

(1) According to the TDA clustering results, the data is divided into several layers (several categories). For each layer, 10% of the datasets in this layer are randomly selected as the labeled sets, and the remaining 90% are unlabeled sets.

(2) Calculating the cosine similarity of unlabeled data with all labeled data, $\cos(\theta) = \frac{\sum_{k=1}^{n} x_{1k}x_{2k}}{\sqrt{\sum_{k=1}^{n} x_{1k}^2}\sqrt{\sum_{k=1}^{n} x_{2k}^2}}$, where x_{1k} is labeled data and x_{2k} is unlabeled data.

(3) Sorting the similarity of the unlabeled sets in ascending order.

(4) Selecting the top m data sets to add to the labeled set, and remove the m data sets from the unlabeled sets.

(5) Repeating steps (2)-(4) until the label sets exceeds $n\%$ of the data sets. Then the several layers of labeled sets and unlabeled sets are respectively combined. The labeled set $(A * n\%)$ is the training set, and the unlabeled set $(A * (100\% - n\%))$ is the testing set, A is the total data.

2.4 Evaluating indicators

In the regression prediction of yield, R^2, Root Mean Square Error ($RMSE$) and Mean Absolute Error (MAE) are selected to measure the regression prediction effect of the model.

(1) R^2, also known as coefficient of determination, reflects the interpretable proportion of the independent variable to the dependent variable. The value range of R^2 is between 0 and 1. The closer R^2 is to 1, the better the fitting effect of the model.

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum\limits_i (y_i - \hat{y}_i)^2}{\sum\limits_i (y_i - \bar{y})^2}, \tag{9}$$

where SST is the sum of squares, and the sum of squares of errors between the original data y_i and the mean value $\bar{y}$ is calculated. SSR is the sum of squares of regression, which calculates the sum of squares of the mean value $\bar{y}$ and the error of fitting data $\hat{y}_i$.

(2) $RMSE$ is the square root of the ratio of the square of the deviation between the observed value $\hat{y}_i$ and the real value y_i and the observation times n. The smaller the value of $RMSE$, the better the regression prediction effect of the mode.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}. \tag{10}$$

(3) MAE is the average of the absolute value of the error between the observed value and the real value. Similarly, it is used to measure the

deviation between the predicted value and the real value. The smaller the MAE value, the better the regression prediction effect of the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{11}$$

3 Results

In this part, TDA and multi-factor analysis of variance are performed to analyze of the chemical reaction conditions that affecting high yield. The convergence, prediction accuracy, and interpretability of the LightGBM model are tested and analyzed. And the stratified diversity sampling strategy is used to enhance the performance of model.

Experimental environment: each experiment is the result of an average of 100 trials with the same configuration. Computer configuration is as follows: Brand: Dell; CPU: Intel(R) Core (TM) i7-7700HQ CPU @2.80GHz(8CPUs), 2.8GHz; Memory type: DDR4.Software: under Python3.7 scikit-learn module or MATLAB R2020a on a 2.80GHz machine with 24.00GB RAM.

3.1 Source data

This paper selects the data published by Ahneman et al. [12] on the Buchwald-Hartwig coupling reaction. Ahneman et al. used an ultra-high-throughput device for coupling reactions and obtained data for 4608 reactions (including controls) spanning different reaction combinations consisting of 4 components, including 23 isoxazole additives, 15 aryl or Heteroaryl halide, 4 palladium catalyst ligands and 3 bases. The yields of these reactions are used as the model output. The effective experimental data are 3960 (Among them, the yields of 5 sets of experimental data are missing, so the 5 sets of experiments are deleted in this paper). Chemical descriptors of reactants, catalysts, and additives involved in the reaction are independent variables, and the corresponding reaction yields are dependent variables. Ahneman et al. to avoid prohibitively time-consuming analysis and logging of computational data, they developed software to submit molecular, atomic, and vibrational property calculations to Spartan and subsequently extract these features from the resulting text files for accessibility to a general user. The program requires only the input of reagent structures in the Spartan graphical user interface and specification of the reaction components in a Python script; it is applicable to any reaction type. The program then generates the data table that can be used for modeling. In total, 120 descriptors are extracted by the software to characterize each reaction.

However, more descriptors may have a large correlation between features, which leads to over-fitting and increases computation time. Therefore, this paper uses the feature descriptor data filtered by our team as the input data for all subsequent algorithms. It has been stated in the text that the obtained 21 descriptors can well replace the original 120 descriptors [36, 37].

Table 1. The 21 feature descriptors extracted through the feature screening method based on importance and correlation.

	feature descriptors		feature descriptors
1	additive_dipole_moment	12	aryl_halide_ovality
2	additive_electronegativity	13	aryl_halide_*C3_NMR_shift
3	aryl_halide_E_HOMO	14	base_dipole_moment
4	aryl_halide_dipole_moment	15	ligand_V10_intensity
5	aryl_halide_electronegativity	16	additive_V1_frequency
6	aryl_halide_molecular_weight	17	additive_V1_intensity
7	additive_*C4_electrostatic_charge	18	ligand_V9_intensity
8	ligand_*C7_electrostatic_charge	19	aryl_halide_V1_frequency
9	base_surface_area	20	aryl_halide_V1_intensity
10	additive_*C3_electrostatic_charge	21	aryl_halide_V3_frequency
11	additive_*C4_NMR_shift		

3.2 Association analysis between reaction conditions and yield

To further explore the internal relationships within the Buchwald-Hartwig coupling reaction data and infer the possible conditions for high-yield reaction, this section analyzes the correlation analysis between reaction conditions and yield based on TDA and multi-factor analysis of variance.

3.2.1 TDA-based association analysis between reaction conditions and yield

In statistics, a quantile is a value that divides a dataset into equal portions based on probability. The meaning of quantile represents the proportion of data subset less than a certain value in the total sample set after a data set is arranged from small to large, which provides a good basis for finding data outliers and observing data distribution. In this paper, the reaction yield was divided into two categories, low-yield rate (Low Yield, less than 0.5 quantile points, which is below the sample median of 28.76173), and high-yield rate (High Yield, more than 0.5 quantile points, which is above the sample median of 28.76173), based on the statistical concept of quantile. Then, TDA cluster analysis is used to provide researchers with corresponding decision-making information.

In TDA, a circle represents a cluster, and its size indicates the number of samples contained. A larger circle indicates that the corresponding cluster contains more samples. The depth of the circle's color represents the average value of the sample labels in the cluster, which in this case is the average yield. The darker the circle's color, the larger the mean label value. When the distance between clusters is smaller, it typically indicates that the internal samples within those clusters are more similar to each other. In Figure 2, it is obvious that K-Means, PCA, t-SNE [40] and UMAP [41, 42] cannot effectively classify the Buchwald-Hartwig coupling reaction data. But TDA can divide it into two classes. As shown in Figure 2(B), the visualization results vary significantly when the parameters are set differently. As a result, two groups of diverse parameters were randomly selected for experimental comparison and analysis in this paper.

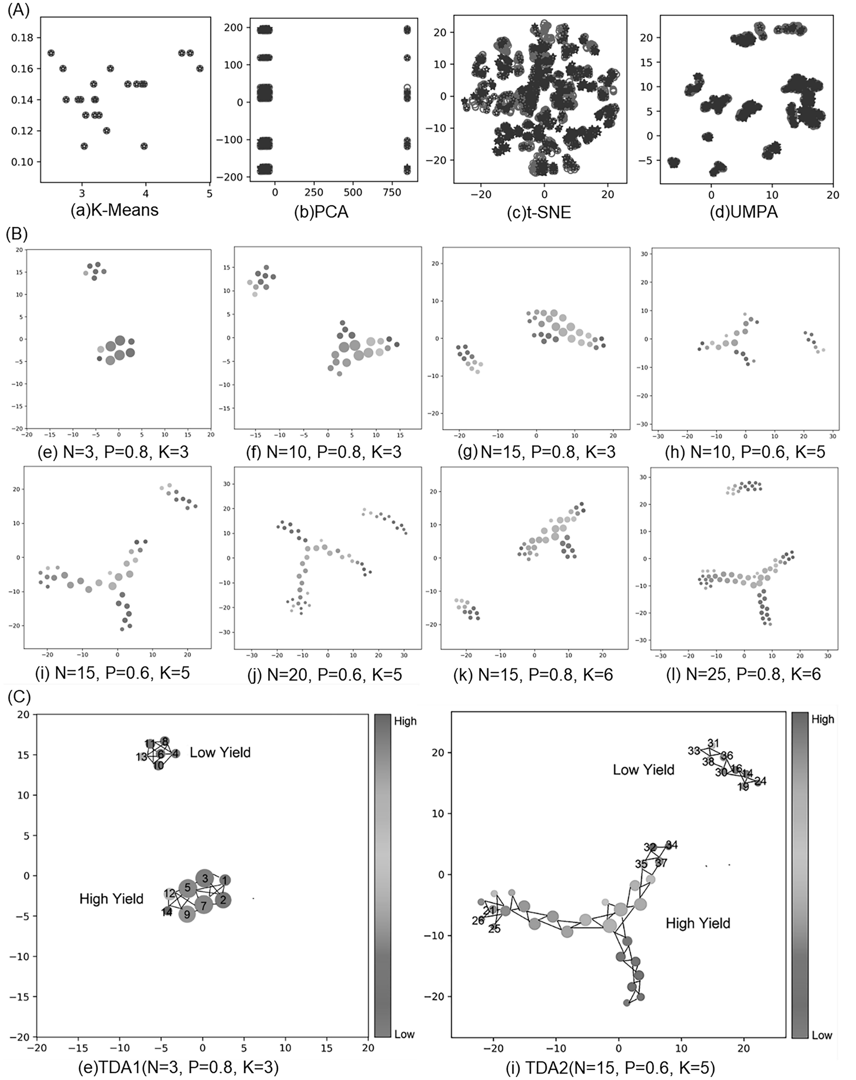


Figure 2. Cluster visualization results. (A) Clustering visualization of K-Means, PCA, t-SNE and UMAP. (B) Visualization of TDA clustering under different parameters. (C) Two groups in (B) are randomly selected for detailed analysis.

Table 2. Detailed analysis of Figure 2(e) and Figure 2(i).

		Cluster	Cluster yield mean	Reaction conditions
(e) TDA1 (N=3, P=0.8, K=3)	Low-yield	Clusters 4,6,8,10,11	< 28.7617	All samples only contained 5th aryl, and 17th additive isn't found in the low-yield samples.
		Cluster 13	> 28.7617	
	High-yield	Cluster 1	< 28.7617	All samples don't contain 5th aryl, and 5th and 7th aryl aren't found in the high-yield samples, which 7th is chloride.
		Clusters 2, 3, 5, 7, 9, 12, 14	> 28.7617	
(i) TDA2 (N=15, P=0.6, K=5)	Low-yield	Clusters 14,16,19, 24,30,36	< 28.7617	All samples only contained 5th aryl, and 17th additive isn't found in the low-yield samples.
		Clusters 31,33,38	> 28.7617	
	High-yield	Upper arm	> 28.7617	All samples don't contain 5th aryl, and 5th and 7th aryl aren't found in the high-yield samples, which 7th is chloride.
		Below arm	< 28.7617	
		clusters 21,25,26	< 28.7617	
		Remaining clusters	> 28.7617	

Through the analysis of each cluster of samples in Figure 2C(e), it is found that, (1) it was discovered that the upper six clusters consist of low-yield sample combinations. Except for the 13th cluster, which has the potential to become a low yield due to its corresponding reaction conditions containing only the 5th aryl, the mean yields of the remaining clusters are less than 28.76173%. Therefore, the 13th cluster was also classified as a low yield. The following eight clusters are high-yield sample combinations. Except for the 1st cluster, the mean yields of the other clusters are greater than 28.76173. However, the 1st cluster is still classified as high-yield because its corresponding reaction conditions have the potential to achieve high-yield output, it does not contain 5th aryl. (2) The mean yield corresponding to 14th cluster is 66.71, which is the highest, among which there are only six samples with low yield. Through analysis, it is found that these six samples are the overlapping samples of 14th cluster with 5th, 7th, 9th, and 12th clusters. The reaction conditions corresponding to these samples all contain only the more active 15th aryl in the reaction conditions, which 15th is iodide.

Through the analysis of each cluster of samples in Figure 2C(i), it is found that the same, (1) The upper nine clusters consist of low-yield sample combinations, except for the 38th, 31st, and 33rd clusters where the mean yields were above 28.76173. However, these clusters were still classified as low yield due to their corresponding reaction conditions, which have the potential to result in low yields since they only contain 5th aryl. The Y-shaped structure is high-yield sample combinations, except for the clusters below the arm of the Y-shaped structure, as well as the 21st, 25th, and 26th clusters. However, the mean yields of all the other clusters are greater than 28.76173. The below arm clusters of the Y-shaped structure and the 21st, 25th, and 26th clusters are still classified as high-yield due to their corresponding reaction conditions having the potential to yield high results, they do not contain 5th aryl. And the samples corresponding to 25th and 26th clusters are all samples from 21th cluster. (2) The 32nd cluster corresponds to the same sample as the 14th cluster in Figure 2C(e). The mean yield corresponding to this cluster is 66.71, which is the highest. Only six samples in this cluster have low yield. Through analysis, it is found that these six samples overlap with 32nd cluster with clusters 32nd, 34th, 35th and 37th. The reaction conditions corresponding to these samples all contain only the more active 15th aryl, which 15th is iodide. (3) The below arm clusters of Y-shaped structure were compared with 24th, 19th, 14th, 16th, 30th, and 36th clusters in the low-yield group, and it is found that in the reaction conditions corresponding to the low-yield samples of these 13clusters, they haven't 2nd, 15th additives and 9th, 15th aryls.

Table 3. Further detailed analysis of Figure 2(i).

		Cluster	Reaction conditions
(i) TDA2 (N=15, P=0.6, K=5)	Low-yield	Clusters 14,16,19, 24,30,36	All samples contained only 5th aryl, and 2th, 6th, 15th, 16th, and 17th additive aren't found in the low-yield samples.
		Clusters 31,33,38	All samples contained only 5th aryl, and 1th, 4th, 7th, 10th, 12th, 13th, 14th, 19th, 20th additive isn't found in the low-yield samples.
	High-yield	Upper arm of the Y	All samples don't contain 5th and 10th aryl, and 5th, 7th and 10th aryl aren't found in the high-yield samples, which 7th and 10th are chloride.
		Below arm of the Y	All samples don't contain 5th, 9th, and 15th aryl, 2th and 3th base, and 4th, 5th, 7th, 9th, and 15th aryl aren't found in the high-yield samples, which 4th and 7th are chloride.

Based on the above analysis, it can be concluded that while the visualization results may vary due to different parameters, the classification results are generally consistent. Further comparative analysis suggests that selecting more active reactants, such as iodide or bromide, can lead to a higher reaction yield (In comparison to chloride, iodide and bromide exhibit greater reactivity, with the order of reactivity being iodide > bromide > chloride [43]). And the 5th aryl should not be selected as much as possible, the additive should be selected as much as possible 1st, 2nd, 4th, 6th, 7th, 10th, 12th, 13th, 14th, 16th, 17th, 19th and 20th, and the base should be selected as much as possible 3rd.

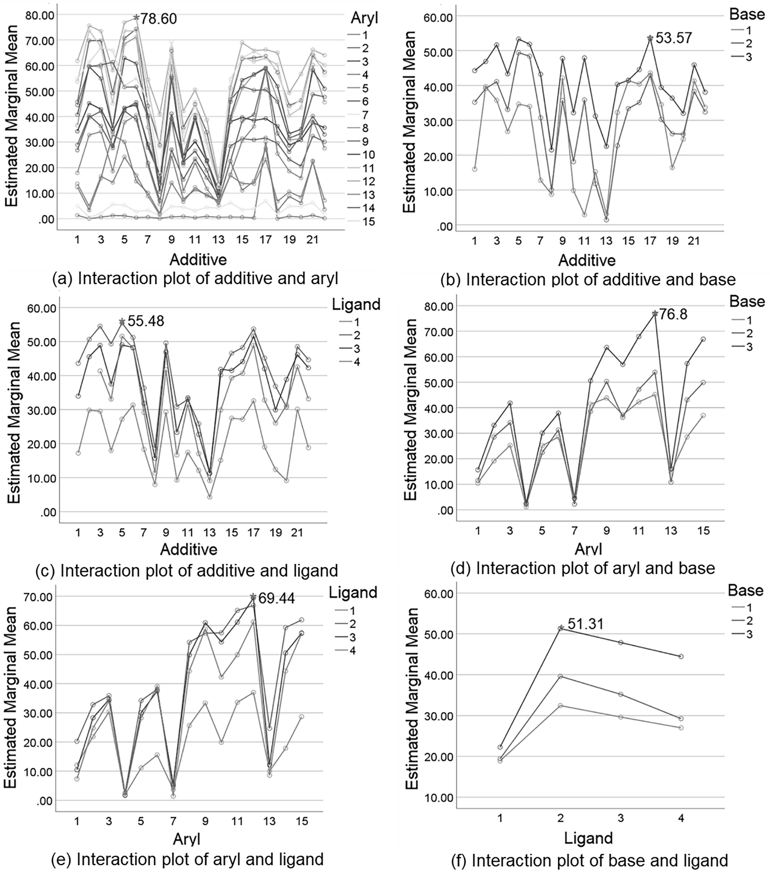
3.2.2 Interaction-based association analysis between reaction conditions and yield

In chemical reactions, the combination of reaction conditions plays a crucial role in determining the outcome. Therefore, it is essential to quantify these interactions and reveal any hidden correlations. This section aims to analyze the effects of various additives, aryls, bases, ligands, and pairwise interactions on yield through multi-factor analysis of variance. And then provides relevant decision-making information for researchers.

To begin with, the pairwise reaction conditions are tested for intersubjective effects. The revised model's values for all six groups of models are less than $\alpha = 0.05$, indicating that the models are statistically significant. The P values of additive, aryl, base, ligand, and their interaction are all less than $\alpha = 0.05$, leading to the rejection of the null hypothesis. It is considered additive, aryl, base, ligand, and their interaction have a significant effect on the yield. (The significance level $\alpha = 0.05$).

In Figure 3(a), it is shown that when changing from the 1st additive to the 22nd additive, choosing a specific aryl can maximize the average yield. For instance, selecting the 12th aryl while using the 1st additive can lead to the highest average yield. Among all combinations of additive and aryl, the combination of the 6th additive and the 12th aryl results in the largest average yield, which is 78.6. Similarly, the Table 4 is got.

According to Table 4, the yield was higher when 5th, 6th and 17th additives, 12th aryl (which 12th is iodide), 2nd and 3rd base, 2nd and 3rd ligand were selected. This observation is consistent with the findings of the TDA clustering analysis discussed earlier.



(a) Interaction plot of additive and aryl

(b) Interaction plot of additive and base

(c) Interaction plot of additive and ligand

(d) Interaction plot of aryl and base

(e) Interaction plot of aryl and ligand

(f) Interaction plot of base and ligand

Figure 3. The pairwise interaction plot of additive, aryl, base, ligand obtained by multi-factor analysis of variance.

Table 4. Pairwise optimal combination of reaction conditions. (In the table, (6, 12) represents 6th additive and 12th aryl are the optimal combination of the additive and aryl.)

	Additive	Aryl	Base	Ligand
Additive	−−	(6,12)	(17,3)	(5,2)
Aryl	(12,6)	−−	(12,3)	(12,3)
Base	(3,17)	(3,12)	−−	(2,3)
Ligand	(2,5)	(2,12)	(3,2)	−−

To sum up, in order to obtain a higher reaction yield, chemically active reactants, such as iodide or bromide should be selected as far as possible, the 5th aryl should be avoided if possible, the 1st, 2nd, 4th, 5th, 6th, 7th, 10th, 12th, 13th, 14th, 16th, 17th, 19th and 20th additives should be selected as much as possible, the 2nd and 3rd bases should be selected as much as possible, and the 2nd and 3rd ligands also should be selected as much as possible.

3.3 LightGBM-based chemical reaction yield prediction

In this section, the convergence and predictive performance of the LightGBM model were investigated. By comparing with ML and deep learning methods, it is proved that the LightGBM not only has good prediction accuracy, but also has faster running speed. Then the stratified diversity sampling was used to enhance the generalization ability of the model.

3.3.1 Parameter optimization and convergence analysis

The optimal parameters of the LightGBM model are typically determined through a combination of cross-validation and grid search training. For example, to find the optimal learning rate, as shown in Figure 4, when the learning rate is 0.12, the highest score is 0.9484.

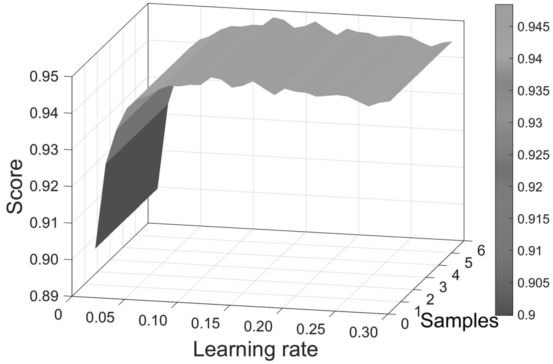


Figure 4. Grid search for learning rate.

Additionally, the convergence of the LightGBM model is analyzed by obtaining its optimal parameters. As depicted in Figure 5, both the training and testing error curves exhibit a downward trend with increasing iterations and eventually stabilize. This indicates that the LightGBM model has converged after training.

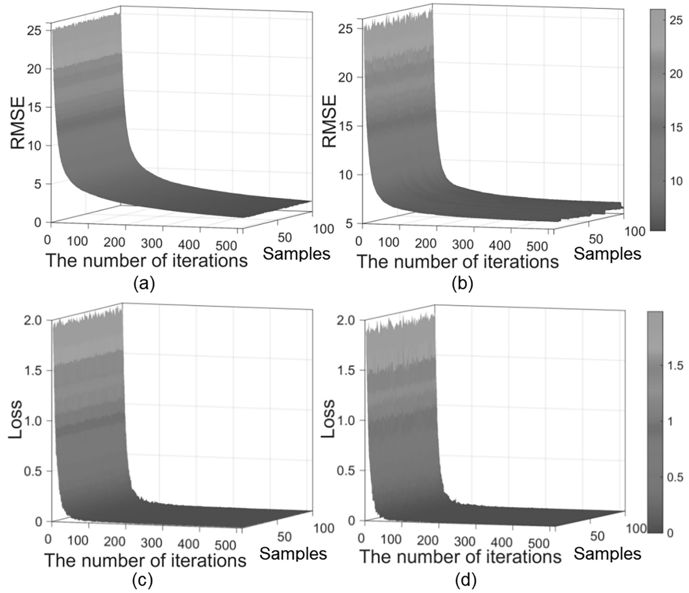


Figure 5. LightGBM's learning curve under10-fold cross-validation. (a)-(b) The RMSE of the training set and testing set varies with the number of iterations. (c)-(d) The absolute error between adjacent iteration steps of the training set and testing set.

3.3.2 Yield prediction accuracy analysis

The regression methods used in this article include MLPR, SVR, AdaBoost, Gradient Boost, Extra Tree, Random Forest, XGBoost and CNN. However, the deep learning models are mainly data-driven and has certain requirements for data. The data screened by our team cannot be applied to the neural network model due to certain limitations. Therefore, when discussing the prediction accuracy of CNN in this section, a large amount of data (source data: 120 feature descriptors) is still selected.

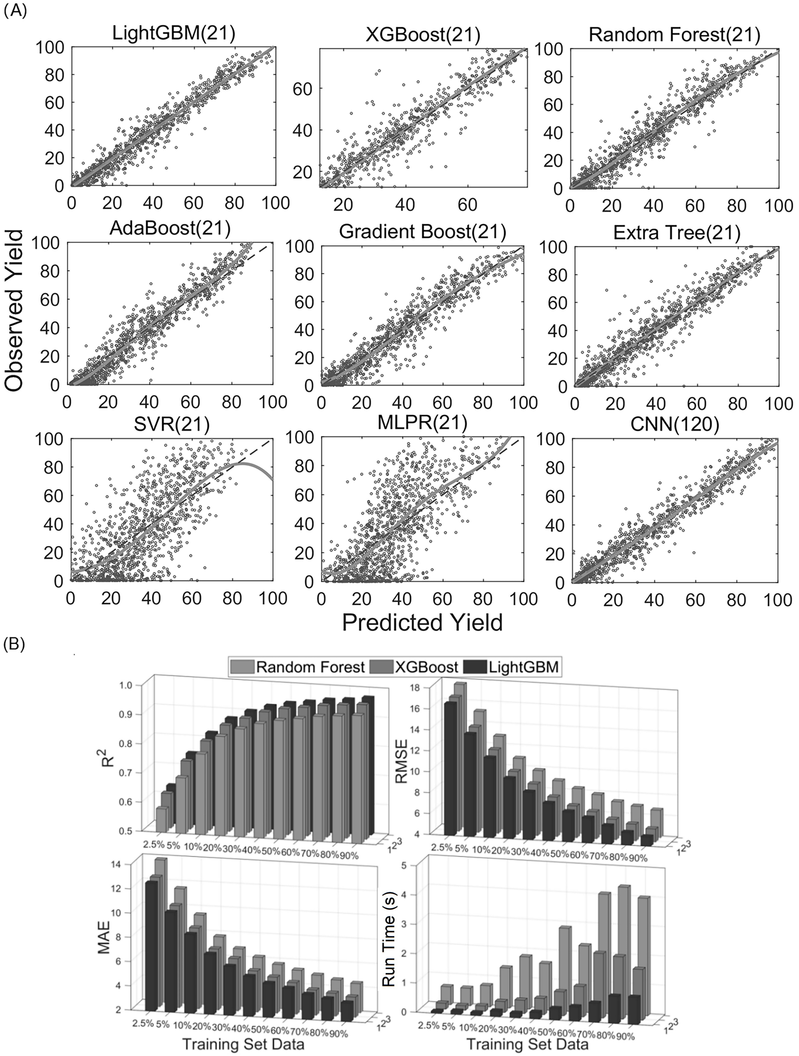


Figure 6. Model prediction performance. (A) Prediction results of different models (training set: testing set = 7:3). (B) LightGBM, XGBoost and Random Forest diagram of different proportions of training data and prediction results.

As shown in Figure 6(A), although general decision tree and ML methods perform nonlinear regression, they are unable to produce relatively accurate prediction results. On the other hand, LightGBM, XGBoost, and CNN have demonstrated better prediction results. However, both XGBoost and CNN require longer training and prediction times, with CNN further demanding a large amount of deep learning data. In contrast, LightGBM performs almost perfectly, achieving an accuracy level of R^2=0.9553, RMSE= 5.7638, MAE=4.0816. Compared with the XGBoost model, the LightGBM model runs more than 3 times faster. This is because LightGBM has introduced two optimization techniques: GOSS and EFB, which allow LightGBM to intelligently select samples and features, reducing computation and memory usage. As a result, it significantly improves training speed and prediction efficiency.

Table 5. Prediction results for nine models. (training set: testing set = 7:3)

Methods	R2	RMSE	MAE	Run Time(s)
LightGBM	**0.9553**	**5.7638**	**4.0816**	**0.6542**
XGBoost	0.9517	5.9733	4.0888	2.2048
Random Forest	0.9295	7.2353	4.9487	4.0732
Extra Tree	0.9240	7.5063	4.8304	5.1285
Gradient Boost	0.9234	7.5386	5.4760	2.5606
Adaboost	0.9201	7.6971	5.9781	4.8961
SVR	0.5309	18.7214	14.9049	41.0222
MLPR	0.5052	19.1672	15.1457	2.8783
CNN	0.9435	6.4801	4.3830	946.4686

For the LightGBM model, it's found that using a significantly smaller subset of the training data of 21 descriptors achieved better predictive power than other methods. As shown in Figure 6(B), with only 5% of the reaction data using LightGBM training to predict the remaining 95% of the reaction data, the result is obvious better than using linear regression prediction results, the accuracy is R^2=0.7446, RMSE=13.7786, MAE=10.2382. When 40% reaction data are used for training prediction, the prediction result of LightGBM can reach the prediction result of Ahneman et al. [12].

In conclusion, the LightGBM algorithm can "learn" enough information from a small amount of data to get better prediction results. It also proves that the 21 descriptors obtained by the integrated feature selection based on importance and relevance can replace the original high-dimensional data.

3.3.3 Stratified diversity sampling-based yield prediction accuracy analysis

In the previous subsection, the predictive ability of LightGBM was verified. And in this subsection, the stratified diversity sampling was used to select training data, which enhancing the performance of the model. Based on the clustering results in the "TDA-based association analysis between reaction conditions and yield" section, the data can be classified into two layers, and diversity sampling can be performed accordingly.

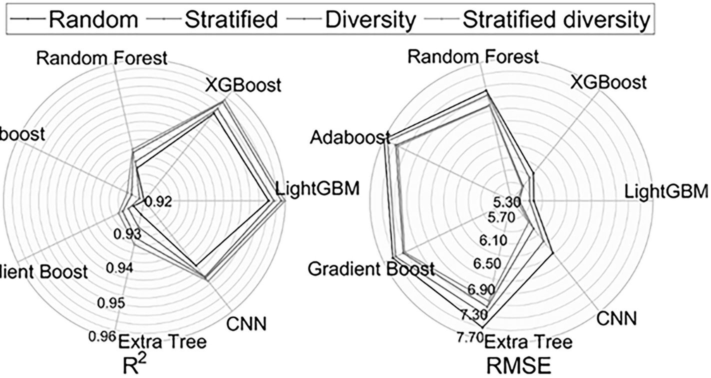


Figure 7. Model prediction results under different sampling methods. (Because the prediction performance of the SVR and MLPR models is poor, it will affect the beauty of the graph, so the SVR and MLPR models are not drawn here, but the prediction results are given in Table 6, training set: testing set = 7:3.)

As shown in Figure 7, for the same model, the training set selected by stratified diversity sampling is better than random sampling. When 70% was selected as training data to predict the remaining 30% of the sample data, using stratified diversity sampling can improve the LightGBM accuracy by 0.0046 for R^2, and reduce by 0.0655 for RMSE. When using 90% of stratified diversity sampling selection as training data to predict the remaining 10% of sample data, the accuracy can even reach R^2=0.9716, RMSE=4.6903.

In a word, compared with the random sampling strategy, the stratified diversity sampling strategy can obtain better prediction performance. And the LightGBM model is superior to other models. It also proves that the TDA clustering results are effective from the side in this paper.

Table 6. Model prediction results under different sampling methods. (Here, we want to compare the prediction results of the four sampling methods under the same model, so the results of MAE and running time are not added, training set: testing set = 7:3.)

Methods	LightGBM		XGBoost		Random Forest	
Index	R2	RMSE	R2	RMSE	R2	RMSE
Sratified diversity	**0.9599**	**5.4924**	**0.9566**	**5.6798**	**0.9350**	**6.9800**
Diversity	0.9589	5.5301	0.9558	5.7044	0.9341	6.9933
Stratified	0.9553	5.7638	0.9534	5.8719	0.9314	7.1568
Random	0.9568	5.6983	0.9517	5.9733	0.9295	7.2353
Methods	Adaboost		Extra tree		Gradient Boost	
Index	R2	RMSE	R2	RMSE	R2	RMSE
Sratified diversity	**0.9254**	**7.4479**	**0.9281**	**7.3227**	**0.9327**	**7.0627**
Diversity	0.9240	7.4871	0.9268	7.3523	0.9308	7.1626
Stratified	0.9215	7.6320	0.9250	7.4868	0.9276	7.2996
Random	0.9201	7.6971	0.9234	7.5386	0.9240	7.5063
Methods	CNN(120)		SVR		MLPR	
Index	R2	RMSE	R2	RMSE	R2	RMSE
Sratified diversity	**0.9491**	**5.846**	**0.536**	**18.5956**	**0.5201**	**18.8675**
Diversity	0.9479	5.9765	0.5343	18.6104	0.5155	18.9024
Stratified	0.9475	6.2369	0.5324	18.6552	0.5107	19.0894
Random	0.9435	6.4801	0.5309	18.7214	0.5052	19.1672

3.3.4 Generalization performance analysis based on out-of-sample predictions and out-of-fold predictions

In order to verify the generalization performance of the proposed method, this paper performs out-of-sample prediction and out-of-fold prediction on the same dataset.

Out-of-sample prediction tests the generalization ability of the model by dividing the data set into two disjoint parts, one to estimate the model and the other to predict. Like [12], isoxazoles in the additive training set (1 to 14 and 16, 17, 20, 23) are used to predict the performance of isoxazoles 15, 18, 19, 21, and 22 in the testing set. The out-of-sample prediction results are shown in Figure 8. Compared with Random Forest-based and XGBoost-based out-of-sample prediction results, LightGBM has a larger R^2, smaller RMSE and MAE, and the shortest Time, which indicates that the LightGBM achieves better out-of- sample prediction effect. It is sufficient to demonstrate that the LightGBM can predict the effect of a new isoxazole or aryl halide structure on the outcome of the Buchwald-Hartwig amination reaction, and identify bases and ligands combinations to provide higher yield.

The concept of out-of-fold prediction is directly related to the concept of out-of-sample prediction. In both cases, predictions are made on samples that were not used during model training, and both allow an estimate of the model's performance when making predictions on new data. The most common method for evaluating a model is to score its predictions during each training session and then average those scores. Another approach is to use out-of-fold prediction, where the predictions for each model are aggregated into a list that summarizes the retained data for each training set as the testing set. Taking the list to get a single accuracy score after all models have been trained. This approach is used to consider that each data appears only once in each testing set. That is, each sample in the training datasets has a prediction during the cross-validation process. After the training process is complete, all predictions can be gathered and compared with the target results to calculate a score. The benefit of this approach is that it can effectively showcase the model's generalization performance.

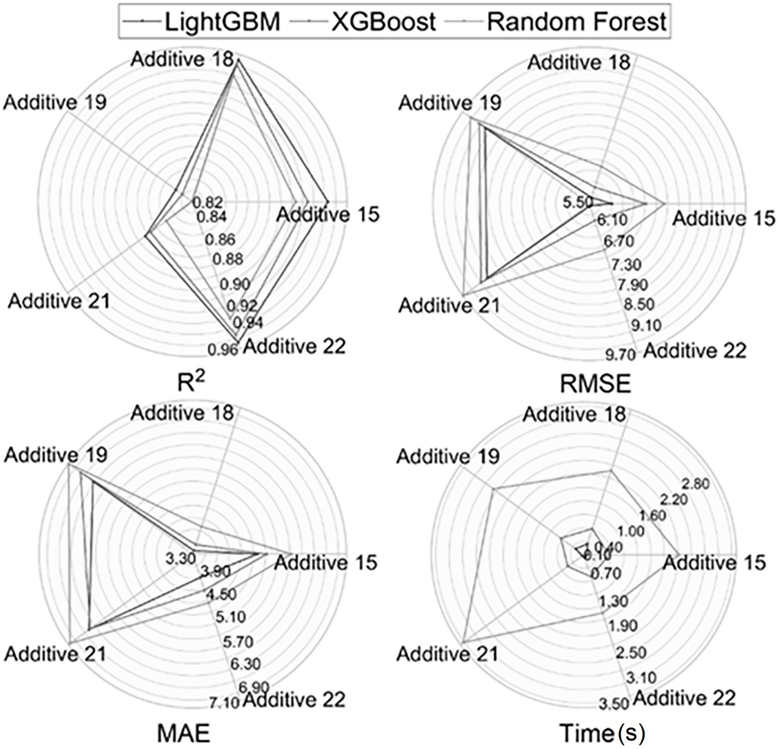


Figure 8. Out-of-sample prediction results.

The out-of-fold prediction results are presented in Figure 9. When compared to the out-of-fold predictions of Random Forest and XGBoost, LightGBM shows a higher R^2 value, smaller $RMSE$ and MAE values, and shorter computation time. These indicates again that the LightGBM achieves better predictions.

The results of out-of-sample prediction and out-of-fold prediction are sufficient to show that the LightGBM can be well used for the prediction of coupled chemical reactions.

Table 7. Out-of-sample prediction results for LightGBM, XGBoost, Random Forest.

	Methods	R2	RMSE	MAE	Run Time(s)
Additive 15	LightGBM	**0.9428**	**6.1124**	**4.8986**	**0.1898**
	XGBoost	0.9246	7.0234	5.1262	0.6692
	Random Forest	0.9136	7.5186	5.7362	2.2215
Additive 18	LightGBM	**0.9538**	**5.7693**	**3.4238**	**0.3559**
	XGBoost	0.9505	5.9671	3.5433	0.7084
	Random Forest	0.9404	6.5508	4.0102	2.0697
Additive 19	LightGBM	**0.8385**	**8.9622**	**6.3477**	**0.3278**
	XGBoost	0.8315	9.1537	6.7293	0.7310
	Random Forest	0.8208	9.4398	7.0925	2.5958
Additive 21	LightGBM	**0.8728**	**8.8927**	**6.4718**	**0.1426**
	XGBoost	0.8669	9.0969	6.4364	0.5400
	Random Forest	0.8492	9.6847	7.0534	3.4195
Additive 22	LightGBM	**0.9540**	**5.5585**	**3.9069**	**0.1768**
	XGBoost	0.9469	5.9724	4.2392	0.6095
	Random Forest	0.9311	6.7990	4.5664	1.4601

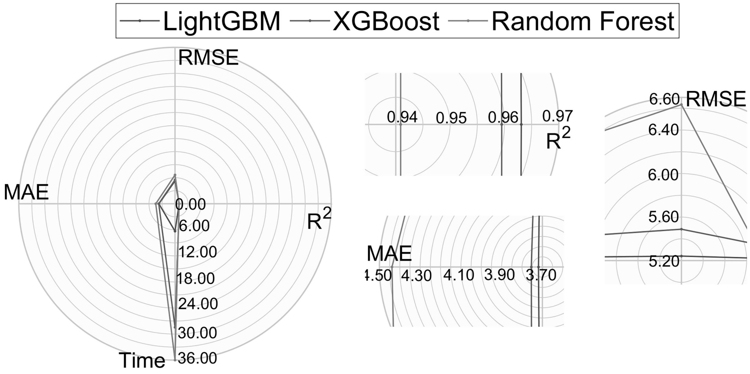
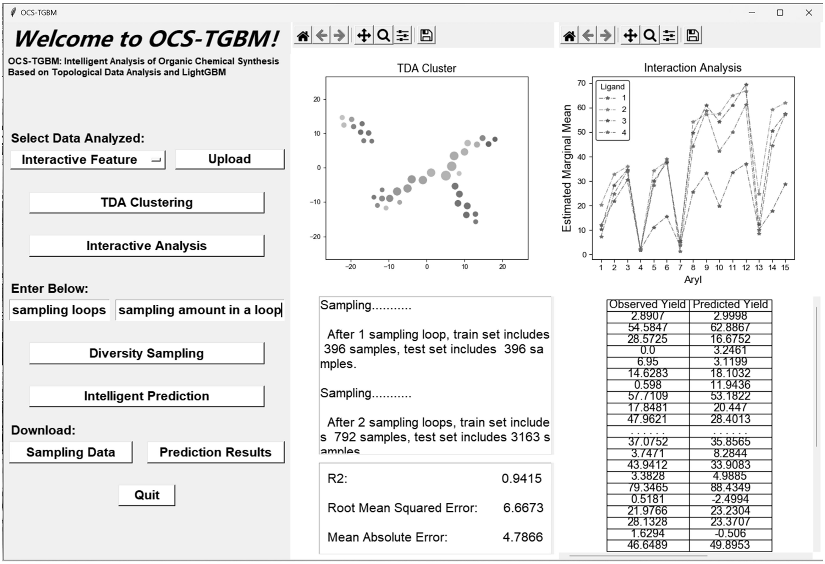


Figure 9. Out-of-fold prediction results. (The right pictures are the partial enlarged picture of R^2, RMSE, MAE.)

Table 8. Out-of-fold prediction results for LightGBM, XGBoost, Random Forest.

Methods	R2	RMSE	MAE	Run Times(s)
LightGBM	**0.9631**	**5.2414**	**3.7192**	**6.3201**
XGBoost	0.9595	5.4885	3.7544	28.4405
Random Forest	0.9409	6.6340	4.4383	35.9912

As shown in Figure 10, for the convenience of users, we has developed a free EXE software called OCS-TGBM, which can implement TDA clustering, multivariate analysis of variance, diversity sampling, and intelligent prediction of reaction yields. All code and software can be found online at https://github.com/cogyh/OCS-TGBM.



```
Sampling...........

 After 1 sampling loop, train set includes
 396 samples, test set includes  396 sa
mples.

Sampling...........

 After 2 sampling loops, train set include
s  792 samples, test set includes 3163 s
amples
```

```
R2:                          0.9415

Root Mean Squared Error:     6.6673

Mean Absolute Error:         4.7866
```

Observed Yield	Predicted Yield
2.8907	2.9998
54.5847	62.8867
28.5725	16.6752
0.0	3.2461
6.95	3.1199
14.6283	18.1032
0.598	11.9436
57.7109	53.1822
17.8481	20.447
47.9621	28.4013
. . . . . .	. . . . . .
37.0752	35.8565
3.7471	8.2844
43.9412	33.9083
3.3828	4.9885
79.3465	88.4349
0.5181	-2.4994
21.9766	23.2304
28.1328	23.3707
1.6294	-0.506
46.6489	49.8953

Figure 10. The system interface of OCS-TGBM.

4 Conclusions

In this paper, the OCS-TGBM model is proposed to explore the internal relationship between reaction conditions and reaction yield in Buchwald-Hartwig coupling reaction, and to make intelligent predictions. The stratified diversity sampling strategy is introduced to improve the model's performance. Finally, an intelligent prediction system with faster training speed, lower memory consumption and better prediction performance is constructed. It provides a new method for researchers to find high-yield reactions, which is helpful to design the required chemical materials more efficiently. Thereby greatly accelerating the process of drug discovery and development.

It is natural to extend the proposed analysis and intelligent prediction system to other chemical reactions beyond coupling reactions. Complementing the advantages of LightGBM and deep neural networks for predicting molecular or drugs design is another interesting and challenging work.

References

[1] P. Ruiz-Castillo, S. L. Buchwald, applications of palladium-catalyzed C-N cross-coupling reactions, *Chem. Rev.* **116** (2016) 12564–12649.

[2] J. F. Hartwig, Evolution of a fourth generation catalyst for the amination and thioetherification of aryl halides, *Acc. Chem. Res.* **41** (2018) 1534–1544.

[3] D. S. Surry, S. L. Buchwald, Biaryl phosphane ligands in palladium-catalyzed amination, *Angew. Chem. Int. Ed.* **47** (2008) 6338–6361.

[4] M. M. Heravi, Z. Kheilkordi, V. Zadsirjan, M. Heydari, M. Malmir, Buchwald-Hartwig reaction: an overview, *J. Org. Chem.* **861** (2018) 17–104.

[5] M. K. Pagels, R. C. Walgama, N. G. Bush, C. Bae, Synthesis of anion conducting polymer electrolyte membranes by Pd-catalyzed Buchwald-Hartwig amination coupling reaction, *Tetrahedron Lett.* **75** (2019) 4150–4155.

[6] X. B. Li, C. X. Zhang, C. C. Wang, W. Q. Ye, Q. Zhang, Z. Y. Z, J. H. Su, Y. F. Chen, H. Tian, Modular synthesis of (C-10 to C-13)-substituted-9, 14-diaryl-9, 14-dihydrodibenzo [a, c] phenazines via a subsequent Buchwald–Hartwig amination and C–H amination strategy, *Chem. Commun.* **56** (2020) 2260–2263.

[7] T. Taeufer, J. Pospech, Palladium-catalyzed synthesis of N, N-dimethylanilines via buchwald-hartwig amination of (hetero)aryl triflates, *J. Org. Chem.* **85** (2020) 7097–7111.

[8] M. Kucharek, A. Danel, Palladium-catalyzed amino group arylation of 1, 3-disubstituted 1 H-pyrazol-5-amine based on Buchwald–Hartwig reaction, *Chem. Heterocyc.* **57** (2021) 633–639.

[9] M. M. Heravi, Z. Kheilkordi, V. Zadsirjan, M. Heydari, M. Malmir, Buchwald-Hartwig reaction: An overview, *J. Org. Chem.* **861** (2018) 17–104.

[10] C. Coley, W. Green, K. Jensen, Machine learning in computer-aided synthesis planning, *Acc. Chem. Res.* **51** (2018) 1281–1289.

[11] S. Szymkuc, E. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. Grzybowski, Computer-assisted synthetic planning: The end of the beginning, *Angew. Chem. Int. Ed.* **55** (2016) 5904–5937.

[12] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, Predicting reaction performance in C-N cross-coupling using machine learning, *Science* **360** (2018) 186–190.

[13] X. C. Mu, J. Dong, L. C. Peng, X. H. Yang, Deep forest-based intelligent yield predicting of buchwald-hartwig coupling reaction, *MATCH Commun. Math. Comput. Chem.* **88** (2022) 5–27.

[14] M. Fujinami, J. Seino, H. Nakai, Quantum chemical reaction prediction method based on machine learning, *Bull. Chem. Soc. Jpn.* **93** (2022) 685–693.

[15] Z. Ahmadvand, M. Bayat, M. A. Zolfigol, Toward prediction of the precatalyst activation mechanism through the cross-coupling reactions: Reduction of Pd (II) to Pd (0) in precatalyst of the type Pd-PEPPSI, *J. Comput. Chem.* **41** (2020) 2296–2309.

[16] W. Yang, T. T. Fidelis, W. H. Sun, Prediction of catalytic activities of bis(imino)pyridine metal complexes by machine learning, *J. Comput. Chem.* **41** (2020) 1064–1067.

[17] K. Yang, Y. Yang, S. Fan, DRONet: effectiveness-driven drug repositioning framework using network embedding and ranking learning, *Brief Bioinf.* **24** (2023) #bbac518.

[18] Z. X. Wu, D. J. Jiang, C. Y. Hsieh, G. Y. Chen, B. Liao, D. S. Cao, T. J. Hou, Hyperbolic relational graph convolution networks plus: a simple but highly efficient QSAR-modeling method, *Brief Bioinf.* **22** (2021) #bbab112.

[19] Z. M. Li, S. C. Zhu, B. Shao, X. X. Zeng, T. Wang, T. Y. Liu, DSN-DDI: an accurate and generalized framework for drug-drug interaction prediction by dual-view representation learning, *Brief Bioinf.* **24** (2023) #bbac597.

[20] Stokes JM, Yang K, Swanson K, A deep learning approach to antibiotic discovery, *Cell* **180** (2020) 688–702.

[21] H. P. Zhang, K. M. Saravanan, Y. Yang, Y. J. Wei, P. Yi, J. Z. H. Zhang, Generating and screening de novo compounds against given targets using ultrafast deep learning models as core components. *Brief Bioinf.* **23** (2022) #bbac226.

[22] S. Ryu, Y. Kwon, W. Y. Kim. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification, *Chem. Sci.* **10** (2019) 8438–8446.

[23] Y. C. Luo, P. P. Wang, M. J. Mou, J. J. Hong, L. Tao, F. Zhu, A novel strategy for designing the magic shotguns for distantly related target pairs, *Brief Bioinf.* **24** (2023) #bbac621.

[24] Y. Y. Yu, X. Wu, Q. Qian, Better utilization of materials' compositions for predicting their properties: Material composition visualization network, *Eng. Appl. Artif. Intel.* **117** (2023) #105539.

[25] M. Druchok, D. Yarish, S. Garkot, Ensembling machine learning models to boost molecular affinity prediction, *Comput. Biol. Chem.* (2021) #107529.

[26] X. Liu, X. J. Wang, J. Wu, K. L. Xia, Hypergraph-based persistent cohomology (HPC) for molecular representations in drug design, *Brief Bioinf.* **22** (2021) #bbaa411.

[27] M. Mishra, H. L. Fei, J. Huan, Computational prediction of toxicity, *Int. J. Data Min.* **8** (2013) 338–348.

[28] H. Zhang, J. Mao, H. Z. Qi, L. Ding, In silico prediction of drug-induced developmental toxicity by using machine learning approaches, *Mol. Div.* **24** (2020) 1281–1290.

[29] S. Moon, S. Chatterjee, P. H. Seeberger, K. Gilmore, Predicting glycosylation stereoselectivity using machine learning, *Chem. Sci.* **12** (2020) 2931–2939.

[30] V. A. Dev, S. Datta, M. R. Eden, M. R. Eden, Hybrid genetic algorithm-decision tree approach for rate constant prediction using structures of reactants and solvent for Diels-Alder reaction, *Comput. Chem. Eng.* **106** (2017) 690–698.

[31] H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification. Proceedings 41st annual symposium on foundations of computer science, *IEEE* (2000) 454–463.

[32] W. J. Beksi, N. Papanikolopoulos, 3D point cloud segmentation using topological persistence, *IEEE* (2016) 5046–5051.

[33] G. Singh, F. Meoli, G. Carlsson, Topological methods for the analysis of high dimensional data sets and 3d object recognition, *PBG@Eurographics* **2** (2007) 91–100.

[34] M. Nicolau, A. J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival, *Proc. Natl. Acad. Sci. USA* **108** (2011) 7265–7270.

[35] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, Extracting insights from the shape of complex data using topology, *Sci. Rep.* **3** (2013) #1236.

[36] L. C. Peng, J. Dong, X. C. Mu, Z. L. Zhang, Y. Q. Zhang, Intelligent predicting reaction performance in multi-dimensional chemical space using quantile regression forest, *MATCH Commun. Math. Comput. Chem.* **87** (2022) 299–318.

[37] J. Dong, L. C. Peng, X. H. Yang, Z. L. Zhang, P. Y. Zhang, XGBoost-based intelligence yield prediction and reaction factors analysis of amination reaction, *J. Comput. Chem.* **43** (2022) 289–302.

[38] T. K. Dey, F. Memoli, Y. S. Wang, Multiscale mapper: Topological summarization via codomain covers, *SLAM* (2016) 997–1013.

[39] G. L. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, LightGBM: A highly efficient gradient boosting decision tree, *NIPS* **30** (2017).

[40] E. Roman-Rangel, S. Marchand-Maillet, Inductive t-SNE via deep learning to visualize multi-label images, *Eng. Appl. Artif. Intel.* **81** (2019) 336–345.

[41] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv.* (2018) **doi:** `https://doi.org/10.48550/arXiv.1802.03426`.

[42] Rahbari A, Rébillat M, Mechbal N, Unsupervised damage clustering in complex aeronautical composite structures monitored by Lamb waves: An inductive approach, *Eng. Appl. Artif. Intel.* **97** (2021) #104099.

[43] T. Yamamoto, M. Nishiyama, Y. Koie,alladium-catalyzed synthesis of triarylamines from aryl halides and diarylamines, *Tetrahedron Lett.* **39** (1998) 2367–2370.