MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **90** (2023) 53–71
ISSN: 0340–6253
doi: 10.46793/match.90-1.053W

CatBoost-Based Framework for Intelligent Prediction and Reaction Condition Analysis of Coupling Reaction

Hengzhe Wang[a], Lichao Peng[b], Li Chang[c], Zixin Li[a], Yanhui Guo[a], Qian Li[d,*], Xiaohui Yang[a,*]

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

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

[c]*Henan Scientific Research platform Service Center, Henan University, Kaifeng, China, 475000*

[d]*Institute of Science and Technology, Henan University, Kaifeng, China, 475000*

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

(Received February 11, 2023)

Abstract

Machine learning is increasingly popular in predicting chemical reaction performance. This study aims to apply the CatBoost algorithm to build an intelligent prediction system for organic chemical reaction yields. The parameter analysis, convergence analysis, prediction accuracy analysis and generalization analysis are carried out. Then, the internal relationship between reaction conditions and yield is excavated through feature importance and SHAP. The results show that the proposed method has the potential as a high-precision tool to assist the optimization of chemical reaction system.

[*] Corresponding author.

1 Introduction

At present, a new round of scientific and technological revolution and industrial transformation are advancing by leaps and bounds. The interdisciplinary integration is developing continuously and emerging technologies represented by information technology and artificial intelligence are developing rapidly. The cross-integration of artificial intelligence algorithms and chemical disciplines, and through theoretical modeling and technological innovation, the research on complex chemical reactions can be carried out with the help of high-speed computer processing capabilities. It is of great significance to carry out intelligent experiments such as synthesis, and prediction to promote the development of academic research. Recently, Artificial intelligence has shown great application potential in the fields of chemical reaction performance [1–5], compound property prediction[6-9], high-performance materials design [10–12], organic synthesis [13, 14].

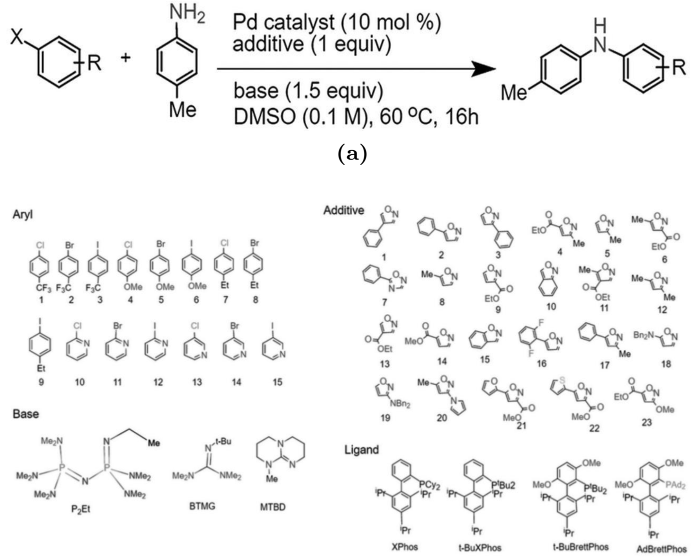
With the help of computer technology, researchers calculate, screen or encode the information in the chemical system in a certain form to form a certain expression of chemical information, i.e., descriptor. The research in the field of chemistry is transformed into the processing of data, which reduces the dependence on personnel to a certain extent. Artificial intelligence algorithms can also mine the potential information inside the large amount of experimental data generated in chemical reaction experiments, helping chemists make reasonable predictions and analysis and greatly improving the efficiency of chemical research and development. A lot of research shows that the artificial intelligence model has high accuracy in solving classification and regression problems. However, among these AI algorithms, deep learning algorithms are like ”black box”, they cannot explain the decision- making process, which is one of the criticisms of using machine learning and artificial intelligence for data research. While they automatically provide useful answers, they do not provide explainable output [15–17]. As a result, we often fail to understand what they are doing and how. Therefore, the researchers considered using an ensemble tree model that is easy to analyze and explain to predict the performance of chemical reactions. In 2018, Ahneman et al. [18] reported the prediction of Buchwald-Hartwig amination yield by Random Forest , which is an advanced study of machine learning methods in the field of multidimensional chemical space prediction. The success of this study has promoted the development of chemical synthesis methods and the study of chemical reaction properties. In ref. [18], the molecular structure is transformed into data descriptors that can be recognized and calculated by computer; The reaction yield data under different reaction conditions were obtained through the high-throughput experimental platform, including different reaction combinations composed of 23 isoxazole additives, 15 aryls and heteroaryl halides, 4 Pd catalyst ligands, and 3 bases. The yields of these reactions are used as the model output. A total of 3960 effective reaction data are generated here. Then, 120 kinds of atomic, molecular and vibration descriptors were obtained through Spartan software calculation and Python script extraction. The Buchwald-Hartwig amination reaction equation and all its reaction components are shown in Figure 1. Finally, using the Random Forest [19] model: 120 descriptors as input, yield as output (70% of reaction data as training set and 30% as test set), the yield was predicted with an accuracy of RMSE = 7.80, R^2= 0.92. However, the work of Ahneman et al. still has the following shortcomings: First, the computational cost of high-dimensional data is high; Second, the Random Forest algorithm has some limitations in the regression prediction: (i) when the number of trees is large, the model training will be slower and time consuming. (ii) If there is noise in the training data of some regression problems, the random forest will overfit.

Therefore, the focus of this paper is to use the feature descriptors corresponding to the reaction conditions after feature screening [20,21], and use the ensemble tree model to build an intelligent prediction system for chemical reaction yield, and explore the internal relationship between reaction conditions and yield deeply. The main contributions are as follows.

(1) A machine learning model called CatBoost [22], an ensemble tree model with superior performance to build a high-accuracy intelligent prediction system for organic chemical reactions.

(2) The internal relationship between reaction conditions and yield can be visualized to provide intelligent assistance for the optimization design

of coupling reaction system.



(a)

(b)

Figure 1. Buchwald-Hartwig amination reaction equation and all its reaction components. (a) A Buchwald-Hartwig amination was used as a model reaction for data generation. Among them: Me, methyl; X, any halide; equiv, equivalent; DMSO, dimethyl sulfoxide; L, ligand; OTf, triflate; i-Pr, isopropyl; R, H or alkyl group; t-Bu, tert-butyl; BTMG, t-butyltetramethylguanidine; MTBD, methyltriazabicyclodecene; Et, ethyl. (b) All reaction components of Buchwald-Hartwig amination reaction.

2 Methods

2.1 CatBoost model

Assume a data set of examples $D = \{(x_k, y_x)\}_{k=1,...,n}$, $x_k = (x_{k1}, ..., x_{km})$ is a random vector of m features, and $y_k \in R$ is a target, which can be either categorical or numerical. Examples (x_k, y_k) are independent and identically distributed according to some unknown distribution $P(,)$. The goal of a learning task is to train a function $F : R^m \rightarrow R$ which minimizes the expected loss $L(F) := EL(y, F(x))$. L is a smooth loss function and

(x, y) is a test example sampled from P independently of training set D. The core of GBDT(Gradient Boosting Decision Tree) [23] is that each tree learns the conclusions and residuals (negative gradients) of all previous trees, and iteratively builds a sequence $F^t : R^m \to R, t = 0, 1, ...$ in a greedy manner, F^t is obtained additively from the previous approximation of F^{t-1}:

$$F^t{=}F^{t-1}{+}\alpha h^t \tag{1}$$

where α is a step size, $h^t : R^m \to R$ is a base predictor to minimize the expected loss:

$$\begin{aligned} h^t &= \underset{h \in H}{\arg\min} L\left(\ \mathrm{F}^{t-1} + \mathrm{h}\right) \\ &= \underset{h \in H}{\arg\min} EL\left(\ \mathrm{F}^{t-1}(\mathrm{x}) + \mathrm{h}(\mathrm{x})\right) \end{aligned} \tag{2}$$

Gradient step h^t is chosen in such a way that $h^t(x)$ approximates $-g^t(x, y)$, where $-g^t(x, y) := \frac{\partial L(y,s)}{\partial s} \left|_{s=F^{t-1}(x)}\right.,$

$$h^t{=}\underset{h \in H}{\arg\min}\, EL(-g^t(x, y) - h(x))^2 \tag{3}$$

In reality, the expectation in (3) is unknown. Generally, usually approximated using the same data set D:

$$h^t{=}\underset{h \in H}{\arg\min}\, \frac{1}{n} EL(-g^t(x_k, y_k) - h(x_k))^2 \tag{4}$$

The chain of shifts is shown below:

(1) the conditional distribution of the gradient $g^t(x_k, y_k)\,|x_k$ is shifted from that distribution on a test example $g^t(x, y)\,|x$

(2) as a result, the base predictor h^t defined by (4) is biased with respect to the solution of Equation (3)

(3) finally, affects the generalization ability of the trained model F_t

The above is the process of model F_t prediction shift. Standard gradient boosting algorithms suffer from some subtle data leakage, which is caused by the iterative fitting method of the model. For a prediction model F that has undergone several enhancement steps, it depends on the target value y of all training examples. In each iteration, the loss function uses the same data set to obtain the gradient of the current model, and then train to get basic learner.

In order to avoid the problem of prediction bias caused by this leakage, CatBoost introduces an "artificial timeline" which can only be calculated using reviously seen example. CatBoost uses the following tricks to handle it: for each example X_k, train a separate model M_k that is never updated using a gradient estimate for this example. With M_k, estimate the gradient on X_k and use this estimate to score the resulting tree. That is, using only the current model trained on previous samples to update the model's gradients on new samples, it provides unbiased residuals and gradients.

Usually in the GBDT framework, the process of building a decision tree can be divided into two stages: selecting the structure of the tree and calculating leaf nodes. In order to choose the best tree structure, in this process, the different splits are enumerated, the tree is then constructed from these splits, the values of the leaf nodes are obtained, the tree is scored, and the best split is selected. In order to enhance the robustness of the model, CatBoost first generates s+1 sequences $\sigma_0, \sigma_1, ..., \sigma_s$ for the training samples, which $\sigma_1, ..., \sigma_s$ are used to construct the decision tree, σ_0 is used to select the value of the leaf nodes, and then use the unbiased estimation of the gradient step size, and then carry out the standard GBDT. In addition, CatBoost uses the symmetric tree [22, 24] as the base predictor. Such trees are balanced and less prone to overfitting. The enhanced forgetting tree has been successfully used in various learning tasks [25]. This particular weak learner significantly speeding up predictions at test time.

2.2 SHAP value

In 2017, Lundberg and Lee proposed the SHAP value as a highly applicable method to explain various models [26]. SHAP explains the predicted value of the model as the sum of the attribution values of each input feature. It [27–29] can reflect the positive and negative influence of the features in each sample, which can combine with the CatBoost model to further explain the relationship between features and predicted results. Assuming that the i-th sample is x_i, the j-th feature of the i-th sample is $x_{i,j}$, the predicted value of the model for the i-th sample is y_i, and the baseline of the whole model (usually the mean value of the target variables of all samples) is y_{base}, then the shake value obeys the following equation:

$$y_i = y_{base} + f(x_{i,1}) + f(x_{i,2}) + ... + f(x_{i,k}) \tag{5}$$

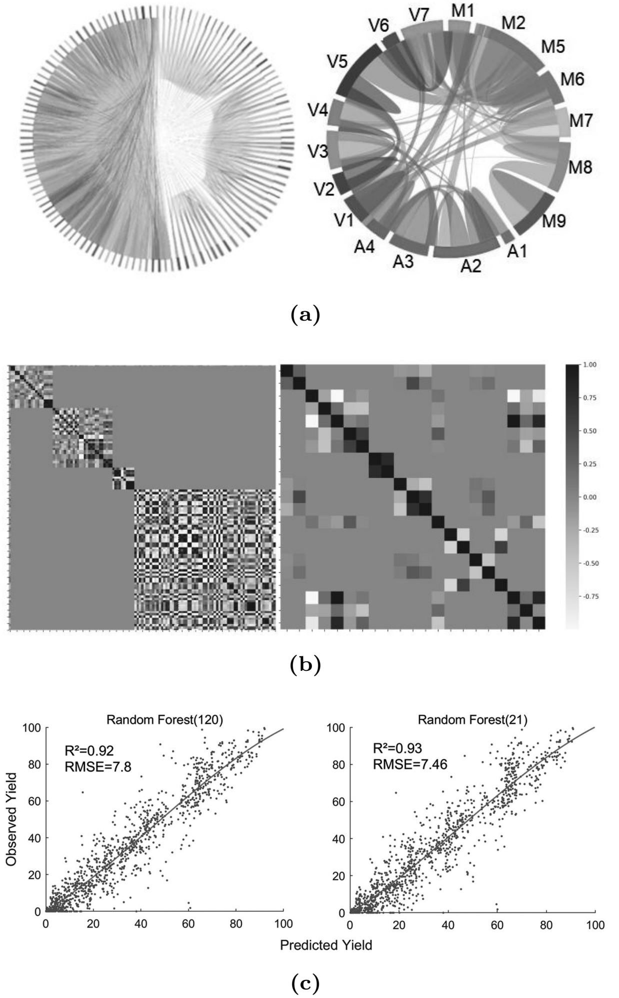
where $f(x_{i,1})$ is the SHAP value of $x_{i,j}$. Intuitively, $f(x_{i,1})$ is the contribution value of the first feature in the i-th sample to the final predicted value y_i. When it shows that $f(x_{i,1}) > 0$, the feature improves the predicted value, it also has a positive effect. On the contrary, it shows that the feature reduces the predicted value, which has a negative effect.

3 Results

The employed machine learning workflow was implemented by the Scikit-learn package (version 0.24.2) in Python (version 3.6.13) or MATLAB 2018a.

3.1 Feature screening analysis

The feature descriptors data [20,21] is the input data of all algorithms later in this study. As shown in Figure 2a,2b, after feature selection, the correlation between descriptors is obviously removed. Then, the prediction accuracy before and after feature screening is verified by Random Forest. As shown in Figure 2c, using the 21 feature descriptors after feature screening for prediction can achieve better prediction accuracy than the original 120 feature descriptors data. It indicates that the 21 descriptors can effectively represent the original descriptor information, which reduce the model complexity and computational cost.



(a)

(b)

(c)

Figure 2. Descriptors filter. (a) The chord diagram of 120 and 21 descriptors (from left to right). (b) The correlation heatmap of 120 and 21 descriptors (from left to right). (c) Random Forest prediction results before and after feature screening

3.2 Model performance analysis

In this subsection, the convergence, predictive performance, and generalization performance of the CatBoost model are analyzed.

3.2.1 Parameter and convergence analysis

Since using the default parameter may not lead to the best performance, its hyper-parameters need to be tuned. The grid search method is an exhaustive search method for specifying parameter values. Combined with the cross-validation method, the parameters of the estimated function are optimized to obtain the optimal parameters. That is, the possible values of each parameter are arranged and combined, and all possible combination results are listed to generate a ”grid”. Each combination is then used for CatBoost training and combined with ten-fold cross-validation to evaluate performance. After the fitting function has tried all parameter combinations, it returns an appropriate learner that automatically adjusts to the best parameter combination.

Although there are many parameters in CatBoost, only a few parameters play a key role in model performance. And a significant advantage of CatBoost is that it does not need to adjust a lot of parameters, just using the default parameters can get good performance. Therefore, we only use the combination of grid search and ten-fold cross-validation to optimize some important parameters, and the other parameters use the default values. Some important parameters are: learning rate, depth, number of trees (iterations), l2-leaf-reg. However, when the parameter search space is large, the grid search method will consume a lot of time and memory, considering this and combining prior experience and historical data, set iterations=400, and select the most likely parameters for the rest of the parameters as the range of grid search.

The learning rate is a very important parameter, so we expand its search interval by the control variable method and combine with ten-fold cross-validation to explore better parameters. As shown in Figure 3a, 0.15 is the best learning rate. To sum up, the CatBoost model parameters are as follows: learning rate is 0.15, depth is 8, iterations is 400, and l2-leaf-reg is 5. The default parameters are used for the rest. After parameter adjustment, the prediction accuracy has been significantly improved.

On this basis, the convergence of the model is analyzed. As shown in Figure 3b, the overall error between adjacent iteration steps decreases with the number of iterations, which indicates that the CatBoost model is convergent.

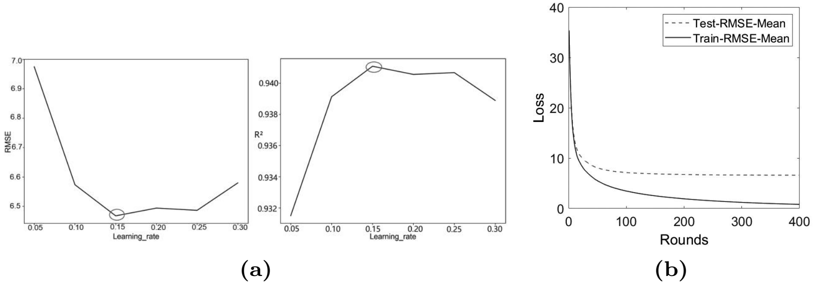


Figure 3. Parameter and convergence result analysis. (a) Learning curve for the learning rate. (b) Error iteration curve obtained by ten-fold cross-validation.

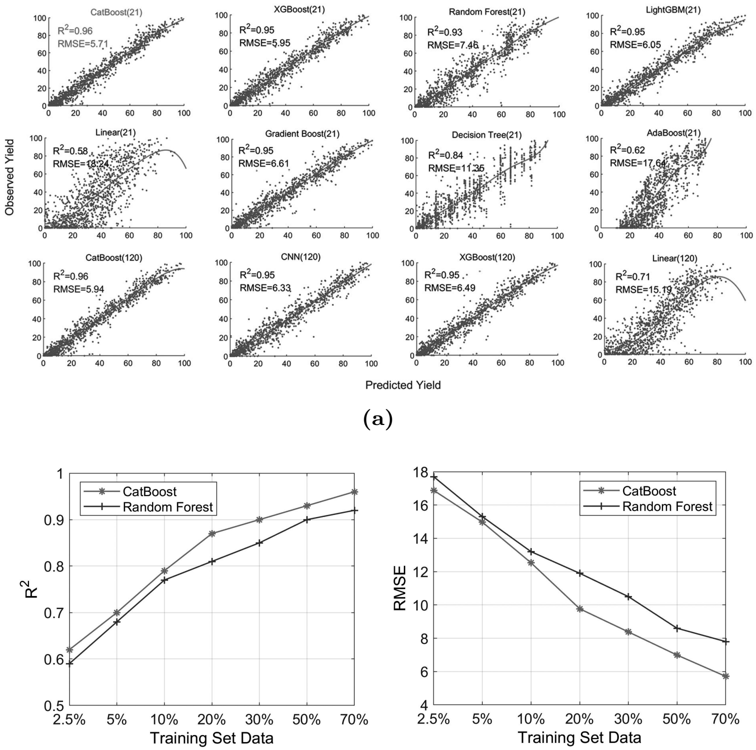
3.2.2 Yield prediction accuracy analysis

With these data in hand, we evaluate the predictive accuracies of linear regression and an array of ML methods using 70% of the data as a training set to predict the remaining 30% (test set). The ML methods used in this section include XGBoost, LightGBM(LGBM), Gradient Boost, Random Forest, Decision Tree, AdaBoost, K-nearest neighbor (KNN), Ridge, Extra tree and the neural network method is convolutional neural networks (CNN).

As shown in Figure 4a, the linear regression method is obviously not suitable for the filtering data. Although decision tree and other machine learning methods have improved, the results are still unsatisfactory. XGBoost and LightGBM have improved significantly, but CatBoost model has better fitting effect on data. CNN has also achieved considerable results, but it is relatively complex and time-consuming. In summary, the CatBoost model is found to be top performer among them, with R^2=0.96, RMSE=5.71.

For the CatBoost model, it is observed that using a significantly smaller subset of the training data of 21 descriptors even achieved good predictive power. As shown in Figure 4b, for the 21 descriptors obtained after filtering, compared with the results of ref. [18], and the prediction results of

CatBoost are all higher than Random Forest. And when training on 50% of the reaction data, the prediction results of CatBoost are already better than the prediction results of the Random Forest in ref. [18]. Using only 10% of the reaction data as a training set to predict the remaining 90% of the reaction data is also better than linear regression using 120 feature descriptors. The above results indicate that (i) CatBoost has no strict requirements on the amount of data, and the prediction performance is still good in small samples, and (ii) the prediction ability of CatBoost will increase with the increase of the number of training subsets.



(a)

(b)

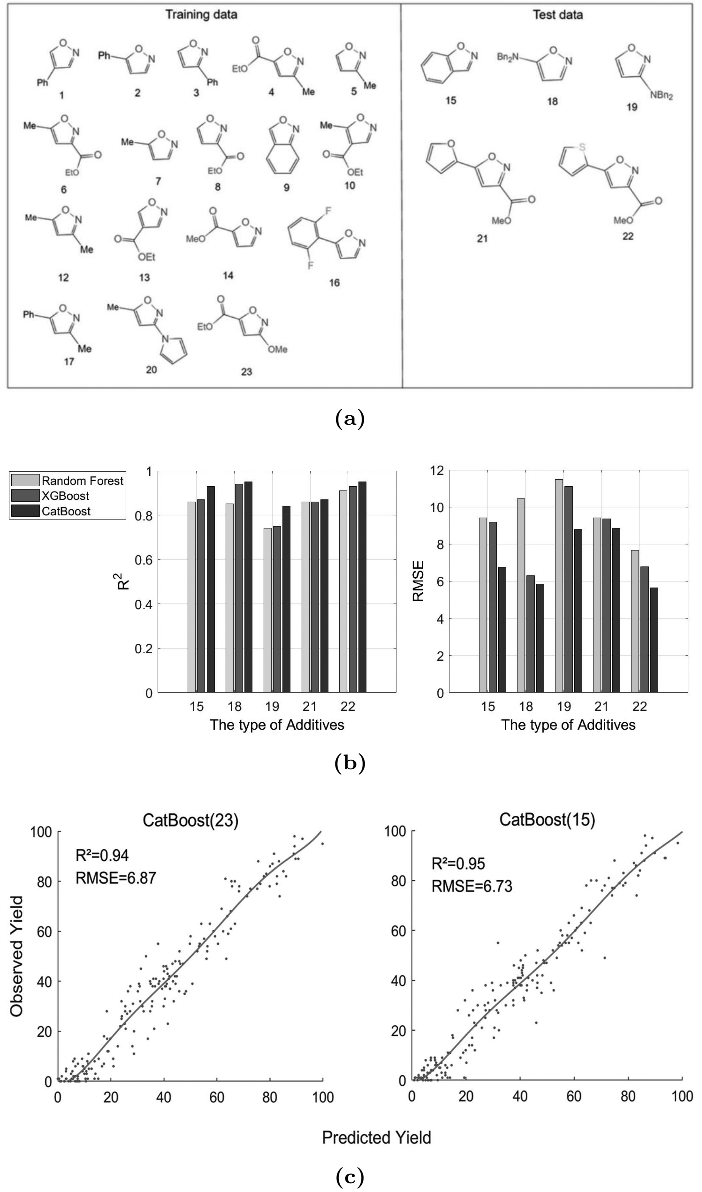
Figure 4. Yield prediction result analysis. (a) Observed versus predicted plots for various ML algorithms and linear regression analysis. (b) The test set performance of CatBoost and Random Forest with sparse data.

3.3 Generalization analysis

In order to verify the generalization performance of the CatBoost model, two experiments are carried out: one is an out-of-sample prediction on the same dataset, and the other is a prediction test on another dataset.

The Out-of-sample prediction tests the generalization of the model by dividing the data set into two disjoint parts, one to estimate the model and the other to predict. Similar to ref. [18], five additives are selected as unknown reaction conditions and the remaining known reaction conditions were used as training data to predict the yield of the unknown reaction conditions. The structure diagram of the out of-sample predicted additive is shown in Figure 5a, and the out of sample prediction results are shown in Figure 5b. The results indicate that (i) compared with the out-of-sample prediction results based on Random Forest and XGBoost, CatBoost has a larger $R^2$and a smaller RMSE, which indicates that our method achieves better out-of-sample prediction effect. (ii) on average, no additive has significant systematic deviation from the prediction of the model, and (iii) our model can predict the effect of a new isoxazole or aryl halide structure on the outcome of the Buchwald-Hartwig amination reaction and determine the combination of bases and ligands to provide the highest yield.

Further, another publicly available Ni-catalyzed cross-coupling reaction data ref. [30] are selected to predict the reaction yield. The coupling reaction data contained 641 reaction samples and 23 descriptors. In ref. [30], Random Forest algorithm was used to predict the yield, and the result was R^2=0.93 and RMSE=7.40, respectively. After using the feature screening method mentioned in the text, we obtain 15 features, and use them as data input. Compared with the original data used in original paper, we can obtain better results with less feature data. Both of the two out-of-sample prediction experiments demonstrate that our model has good generalization ability.



(a)

(b)

(c)

Figure 5. Model generalization analysis. (a) Isoxazoles in the additive training set (1 to 14 and 16, 17, 20, 23) were used to predict the performance of isoxazoles 15, 18, 19, 21, and 22 in the test set. Ph, phenyl; Bn, benzyl. (b) Comparison of out of sample prediction results of Random Forest, XGBoost and CatBoost. (c) Application of our method to other coupling reaction reactions

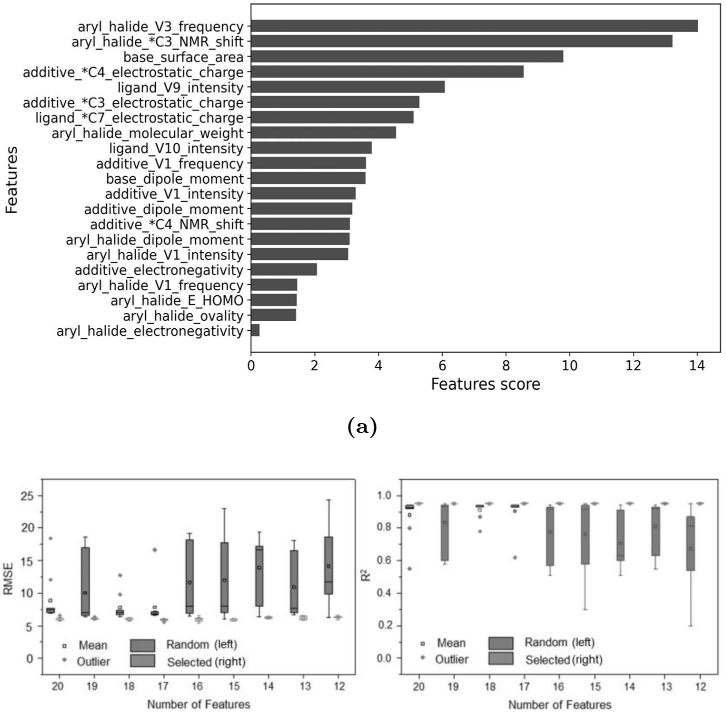
3.4 Explainability analysis

The explainability of the model is as important as the prediction accuracy. In this section, we analyze the importance of features in the modeling process through the importance ranking of Catboost model output, and verify the effectiveness of features with the importance ranking. Further, the correlation between feature descriptors and reaction yields is analyzed by SHAP value analysis.

3.4.1 Feature importance analysis

In order to understand the key features in the model prediction process, we carry out a feature importance analysis. As shown in Figure 6a, we find that three of the five most important descriptors in predicting reaction outcomes are the additive_*C4_electrostatic_charge(* indicates a shared atom), additive_*C3_ electrostatic_charge, and aryl_halide_*C3_NMR_shift. Among the top 10 descriptors in importance ranking, there are three aryl halide descriptors and three additive descriptors. These descriptors suggest that the propensity of additives [31, 32] and halides [33] to act as electrophiles may influence reaction outcomes.

To validate the effectiveness of the features, we sampling 70% as training set, and the top 20-12 descriptors are selected, which are based on the feature ranking (Figure 6a) from high to low, as features to retrain CatBoost. The sampling is repeated for ten times generate ten results used to plot with corresponding feature numbers. Meanwhile, the same procedure is applied on the same number of descriptors randomly sampled from the original 120 descriptors to plot precision as the contrast. As shown in Figure 6b, with the decreased numbers of features from 20 to 12, for the selected features keep stable. In contrast, the results are more volatile with the decrease in the number of randomly selected features. Hence, the feature importance analysis by CatBoost is meaningful.



(a)

(b)

Figure 6. Feature importance. (a) Descriptor importance of trained CatBoost model. And * indicate a shared atom. E, energy; HOMO, highest occupied molecular orbital; V, vibration. (The degree of influence of the change of the feature value on the predicted average value, the more important the feature, the greater the impact). (b) Verification of the feature importance ranking.

3.4.2 Correlation analysis between reaction condition and yield

Knowing the rank of feature importance in model predictions is not enough. It is necessary to understand the correlation effect of feature descriptors on yield during model calculation. So we use the SHAP value to achieve this aim.

As shown in Figure 7, each row represents a feature with a SHAP value, a point represents a sample and a wide area indicates a large number of samples are gathered. The color indicates the feature value. The darker the color is, the larger the feature value is, and the lighter the color is, the smaller the feature value is. It is observed that there is basically a positive correlation between aryl_halide_*C3_NMR_shift and the reaction yield, the higher the value of the descriptor is, the higher yield is. And there is a negative correlation between aryl_halide_V3_frequency and reaction yield, and the higher the value of the descriptor is, the lower yield is.

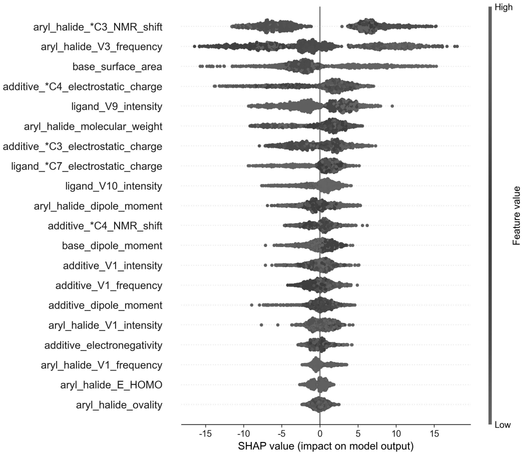


Figure 7. SHAP value analysis.

4 Conclusions

In this paper, an intelligent prediction system with good prediction performance is constructed based on CatBoost algorithm. Our work optimizes the reaction system through machine learning method, realizes automatic and intelligent prediction of reaction yield, improves the credibility of the model, and will help design the required chemical materials more efficiently. In the future, we hope to be able to combine CatBoost with deep neural networks to predict the synthesis of molecules or materials.

***Acknowledgment*:** The original data used in this study is the Buchwald Hartwig amination reaction database provided by Ahneman et al. (https://github.com/Ahnemanlab/rxnpredict).

References

[1] E. M. Gale, D. J. Durand, Improving reaction prediction, *Nature* **12** (2020) 509–510.

[2] M. R. Maser, A. Y. Cui, S. Ryou, T. J. DeLano, Y. Yue, S. E. Reisman, Multilabel classification models for the prediction of cross-coupling reaction conditions, *J. Chem. Inf. Model.* **61** (2021) 156–166.

[3] M. H. S. Segler, M. P. Waller, Neural–symbolic machine learning for retrosynthesis and reaction prediction, *J. Chem. Eur.* **23** (2017) 5966–5971.

[4] M. Haghighatlari, J. Li, F. Heidar–Zadeh, Y. C. Liu, X. Y. Guan, T. Head–Gordon, Learning to make chemical predictions: The interplay of feature representation, data, and machine learning methods, *Chem* **6** (2020) 1527–1542.

[5] P. Schwaller, B. Hoover, J. L. Reymond, H. Strobelt, T. Laino, Extraction of organic chemistry grammar from unsupervised learning of chemical reactions, *Sci. Adv.* **7** (2021) #eabe4166.

[6] C. Feldmann, M. Philipps, J. Bajorath, Explainable machine learning predictions of dual-target compounds reveal characteristic structural features, *Nature* **7** (2021) 21594.

[7] X. F. Wang, Z. Li, J. M. Jiang, S. Wang, S. G. Zhuang, Z. Wei, Molecule property prediction based on spatial graph embedding, *J. Chem. Inf. Model.* **59** (2019) 3817–3828.

[8] S. Boobier, D. R. J. Hose, A. J. Blacker, B. N. Nguyen, Machine learning with physicochemical relationships: solubility prediction in organic solvents and water, *Nature* **11** (2020) 5753.

[9] J. G. Cumming, A. M. Davis, S. Muresan, M. Haeberlein, H. M. Chen, Chemical predictive modelling to improve compound quality, *Nature Rev. Drug Discov.* **12** (2013) 948–962.

[10] Y. Wu, J. Guo, R. Sun, Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells, *Nature* **6** (2020) 120.

[11] A. Challapalli, G. Li, Machine learning assisted design of new lattice core for sandwich structures with superior load carrying capacity, *Nature* **11** (2021) 18552.

[12] W. Sun, Y. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li, K. Sun, Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials, *Sci. Adv.* **11** (2019) #eaay4275.

[13] J. M. Granda, L. Donina, V. Dragone, De-Liang Long, L. Cronin, Controlling an organic synthesis robot with machine learning to search for new reactivity, *Nature* **559** (2018) 377–381.

[14] A. F. de. Almeida, R. Moreira, T. Rodrigues, Synthetic organic chemistry driven by artificial intelligence, *Nature* **3** (2019) 589–604.

[15] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Mach. Intell.* **1** (2019) 206–215.

[16] R. Dybowski, Interpretable machine learning as a tool for scientific discovery in chemistry, *New J. Chem.* **44** (2020) 20914–20920.

[17] J. Feng, J. L. Lansford, M. A. Katsoulakis, D. G. Vlachos, Explainable and trustworthy artificial intelligence for correctable modeling in chemical sciences, *Sci. Adv.* **6** (2020) 3204–3218.

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

[19] L. Breiman, Random Forests, *Mach. Learn.* **1** (2001) 5–32.

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

[21] 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.

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, *Comput. Sci.* **5** (2018) 6637–6647.

[23] J. H. F. Ann, Greedy function approximation: a gradient boosting machine, *Ann. Stat.* **5** (2001) 1189–1232.

[24] R. Kohavi, C. H. Li, Oblivious decision trees, graphs, and top-down pruning, *IJCAI* **2** (1995) 1071–1077.

[25] A. Gulin, L. Kuralenok, D. Pavlov, Winning the transfer learning track of yahoo!'s learning to rank challenge with yetirank, *Yahoo! Learning to Rank Challenge.* **14** (2011) 63–76.

[26] S. M. Lundberg, S. L. Lee, A unified approach to explaining model predictions, *NIPS* **31** (2017) 4768–4777.

[27] L. S. Shapley, A value for n-person games, *Contrib. Theory Games* **2** (1953) 307–317.

[28] M. T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you?: Explaining the predictions of any classifier, *ACM* **3** (2016) 1135–1144.

[29] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, *Knowledge Inf. Sys.* **3** (2014) 647–665.

[30] K. Wu, A. G. Ahneman. Parameterization of phosphine ligands demonstrates enhancement of nickel catalysis via remote steric effects, *Nature Chem.* **9** (2017) 779–784.

[31] Y. Fall, C. Reynaud, H. Doucet, M. Santelli, Ligand-free-palladium-catalyzed direct 4-arylation of isoxazoles using aryl bromides, *Eur. J. Org. Chem.* **24** (2009) 4041–4050.

[32] M. Shigenobu, K. Takenaka, H. Sasai, Palladium-catalyzed direct C–H arylation of isoxazoles at the 5-position, *Angew. Chem. Int. Ed. Engl.* **54** (2015) 95729576.

[33] W. Zhang, L. X. Lu, W. Zhang, Y. Wang, S. D. Ware, J. Mondragon, J. Rein, N. Strotman, D. Lehnherr, K. A. See, S. Lin, Electrochemically driven cross-electrophile coupling of alkyl halides, *Nature* **604** (2022) 292–297.