MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **88** (2022) 5-27
ISSN: 0340-6253
doi: 10.46793/match.88-1.005M

Deep Forest–Based Intelligent Yield Predicting of Buchwald-Hartwig Coupling Reaction

Xuechun Mu[1], Jing Dong[1], Lichao Peng[2]*, Xiaohui Yang[1]*

[1]*Henan Engineering Research Center for Artificial Intelligence Theory & Algorithms, School of Mathematics and Statistics, Henan University, Kaifeng, China, 475004*
[2]*National & Local Joint Engineering Research Center for Applied Technology of Hybrid Nanomaterials, Henan University, Kaifeng, China, 475004*

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

(Received November 22, 2021)

Abstract

Buchwald-Hartwig coupling reaction is widely used in organic chemical synthesis, yield prediction is particularly important. In 2018, ***Science*** reported a yield prediction method based on random forest, but this method lacks feature learning. Therefore, an intelligent prediction and analysis method of coupling reaction yield based on deep forest is proposed. Combined with the advantages of deep learning and ensemble learning, the new deep model in the form of non-neural network is explored, which has good characterization learning ability and low difficulty in adjusting parameters, realizes the efficient prediction of chemical reaction, and analyzes the factors that have a significant impact on the prediction of reaction yield.

1 Introduction

With the development of organic technology, more and more experts and scholars began to apply it to drug synthesis. In the process of drug research and development, new molecular

compounds can be generated by the combination reaction of different compounds. Organic synthesis plays an innovative role in the research and development of new drugs and the optimization of old drugs.

Functionalized aromatic heterocyclic amines are important structures for the synthesis of drug molecules. Chemical researchers hope to find other ways to synthesize C-N bonds with mild reaction conditions and high efficiency. The early methods used to construct C-N bonds mainly include Jourdan reaction [1], Ullmann reaction [2], Goldberg reaction [3], but these reactions are not suitable for large-scale production and life because of their harsh conditions. Buchwald Hartwig coupling reaction is a direct and effective method to construct C-N bond.

In 1994, John F. Hartwig [4] and Stephen L. Buchwald [5] found palladium catalyzed coupling reaction between aryl bromide and amine without the participation of organotin compounds. In 2010, Richard F. heck, EI Ichi Negishi and Akira Suzuki were awarded the Nobel Prize in Chemistry for developing "palladium catalyzed cross coupling method in organic synthesis". This technology reduces costs and protects the environment. It has been widely used in global scientific research, pharmaceutical development and production, luminescent materials and electronic industry materials. Subsequently, Buchwald Hartwig coupling reaction developed continuously to obtain high-yield amination products in a cleaner and efficient way [6-10].

Scholars in the field of chemistry mainly change the reaction products in Buchwald Hartwig coupling reaction to achieve better prediction results. With the development of computer technology and artificial intelligence, machine learning (ML) began to be used by chemists to assist in processing chemical data. ML overcomes the difficulties of high characteristic dimension of chemical reaction, toxicity of reactants, complex reaction process and difficult prediction. It shows more and more competitiveness in the research of chemical reaction prediction [11-12], drug performance prediction [13-16], screening target compounds [17-18], and molecular material design [19-20]. In 2018, Doyle [21] and others proposed a method to predict the yield of Buchwald Hartwig coupling reaction based on random forest algorithm [30] in ***Science***, and predicted Buchwald-Hartwig coupling reaction with high accuracy with a goodness of fit of 0.92 and root mean square error of 7.8. It is proved that the ML method can use the data obtained through high-throughput experiments to predict the

performance of multi-dimensional chemical space synthesis reactions. Based on the research of Doyle et al., Peng [22] proposed a quantile regression forest probability density prediction model in 2021. By training the data after feature selection, the prediction interval of Buchwald-Hartwig reaction yield under different quantiles is obtained. Doyle's point prediction is extended to interval prediction, and the value range of the estimated value is judged at a certain probability level.

However, both random forest and quantile regression forest have some defects in dealing with characteristics. Quantile regression forest will select features before training the model, which will cause irreversible loss of information. Both methods do not consider feature learning, and the information is not fully utilized. Aiming at the problem that traditional machine learning lacks feature learning in prediction, an intelligent prediction method of coupling reaction yield based on deep forest is proposed. The model is improved on the database of Buchwald-Hartwig coupling reaction (Fig. 1) obtained by Doyle et al. Compared with the previous work of Doyle and our team, this paper combines deep learning and statistical test indicators to realize the intelligent prediction and analysis of chemical yield with higher accuracy by making full use of characteristic information.

The flow chart is shown in Figure 2. Firstly, the three feature descriptors (molecular descriptor, atomic descriptor, and vibration descriptor) calculated by Spartan from the high-throughput screened Buchwald Hartwig coupling reaction results and the corresponding yield categories of different descriptor ratios are introduced into the deep forest model. Then, the features are trained layer by-layer-by using each layer of cascade forest, Finally, the predicted yield category can be obtained. In addition, by calculating the importance of characteristic descriptors, the significant factors affecting the yield of coupling reaction can be analyzed.

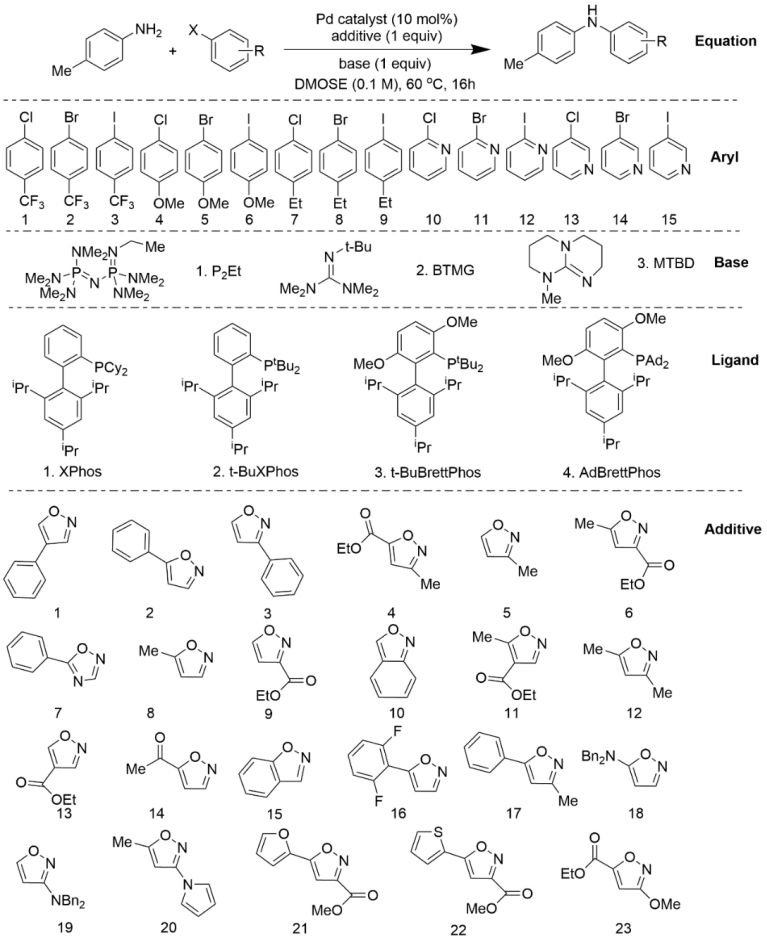

Figure 1. All reaction components of Buchwald Hartwig coupling reaction.

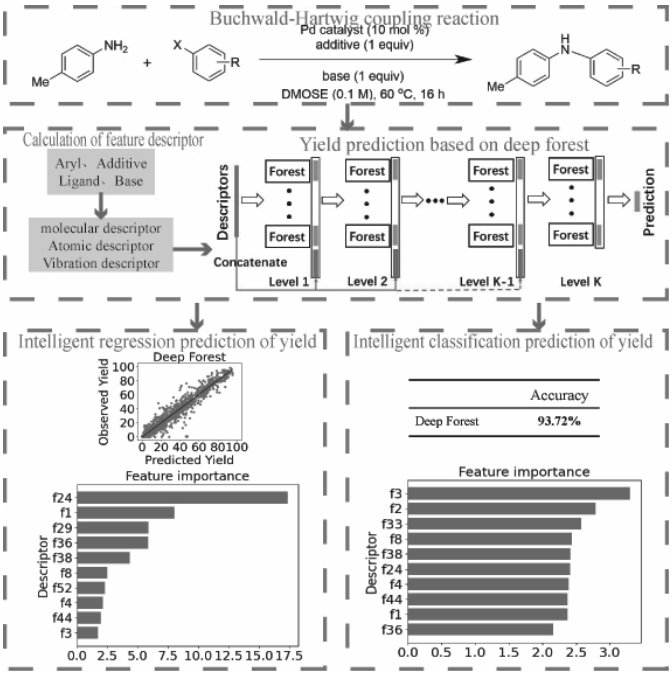

Figure 2. Flow chart of intelligent prediction and analysis of chemical reaction yield.

2 Deep Forest

Deep forest [23] is a new non-neural network deep model. The deep forest algorithm skillfully combines the idea of deep learning and ensemble learning. By integrating the random forest composed of decision tree, it increases the complexity of the model, and trains the features layer-by-layer, to achieve the purpose of representation learning by the base learner, which can effectively improve the effect of classification or regression. At the same time, compared with general deep learning methods, such as Convolutional Neural Networks (CNN), deep forest has fewer hyperparameters, better robustness to hyperparameters and better performance on small sample data sets. In addition, deep forest can be calculated in parallel. The time required to run deep forest on the CPU is like that to run deep neural network with GPU, so deep forest is more practical.

2.1 Deep classified forest

The classification and prediction process of deep forest is shown in Figure 3. The original features and corresponding categories are input into the deep forest model for training. After the first-level training, the class probability vector obtained from each random forest and the original features are joined together as the input of the next level. When the prediction accuracy of the cascaded forest at the next level does not significantly improve compared with the previous level or reaches the maximum upper limit set, the model stops training and takes the average value of the class probability output by each random forest in the last level of the cascade, and the category to which the largest class probability belongs is the final prediction category.

In the deep forest model, there is a T-level cascade, and each cascade is composed of L forests. The training sample input in the t-level cascade is $(x^t, y), t = 0,1,\ldots,T$. Where, x^t represents the feature vectors of the training samples input in the T-level linkage, and y represents the category corresponding to each feature vector. The input feature x^t received in the t-level cascade is the splicing of the original feature x^0 and the output x^{t-1} in the $t-1$ cascade, so the combined feature is expressed as,

$$x^t = \left(p_{t,1}\left(x^{t-1}\right),\ldots,p_{t,L}\left(x^{t-1}\right),x^0\right), \tag{1}$$

where, $p_{t,L}(x)$ represents the probabilistic vector of feature x obtained through the L-th forest in the t-level cascade training.

The final quasi-probability vector is the mean value of all forest prediction probabilities in the last cascade,

$$p(x) = \frac{\sum_{l=1}^{L} p_{T,l}\left(x^{T-1}\right)}{L}, l = 1,2,\ldots,L. \tag{2}$$

Suppose there are class c of training samples, $p(x) = (p_1(x), p_2(x),\ldots,p_c(x))$, the category corresponding to the largest probability in the class probability vector is predicted to belong to the category: $\arg\max_c p(x)$.

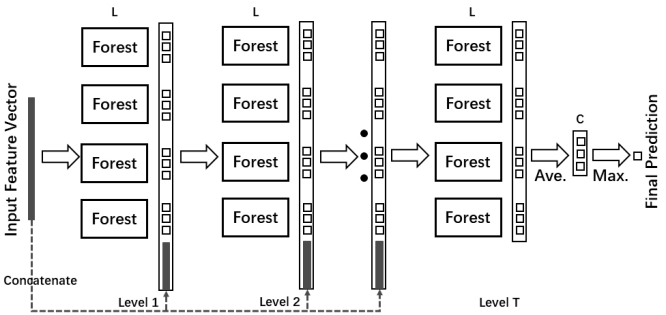

Figure 3. Flow chart of deep forest classification. The class probability vector output by the random forest in each cascade is concatenated with the original feature, which serves as the input of the next hierarchy and outputs the class probability until the last layer. The category corresponding to the maximum value after averaging is the final prediction category.

The algorithm flow of predicting chemical yield category based on deep forest is shown in algorithm 1.

Algorithm 1: The intelligent Predicting Reaction Performance in Multi-Dimensional Chemical Space Using Deep classified Forest algorithm.

Input：Training set $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, $x_i\in\Re^m$, $y_i\in\{1,\ldots,C\}$.

Where x_i represents the feature descriptor and y_i represents the corresponding yield category.

1：**for** $t=1,2,\ldots,T$ **do**

2： **for** $l=1,2,\ldots,L$ **do**

3： According to the deep forest algorithm [35], the L-th random forest of the t-th cascade is trained.

4： Calculate the class probability vector $p_{t,l}(x^t)$ of the L-th random forest in the t-th cascade.

5： $x^{t+1}=(p_{t,1}(x^t),\ldots,p_{t,l}(x^t),x^0)$.

6： **end for**

7： Use x^{t+1} for the input of the next cascade.

8：**end for**

9：Class probability vector from deep forest prediction $p(x)=\frac{\sum_{l=1}^{L} p_{T,l}\left(x^{T-1}\right)}{L}$.

Output：Category of yield $=\arg\max_{c} p(x)$.

2.2 Deep regression forest

The process of regression prediction for deep forest is shown in Figure 4. The original features and corresponding real values are input into the deep forest model for training. After the first-level training, the predicted values and original features obtained from each random forest are joined together as the input for the next level of training. When the Mean Square Error (MSE) of the cascade forest of the next layer does not significantly improve compared with the previous layer or reaches the set maximum upper limit layer, the model stops training and outputs the final predicted yield. Finally, the predicted value obtained by each random forest in the last hierarchy is averaged to obtain the final predicted value.

In the deep forest model, there is a K-level cascade, and each cascade is composed of L forests. The training sample input in the k-level cascade is $\left(x^{k}, y\right), k=0,1,\ldots,K,$ where x^{k} represents the feature vectors of the training samples input in the K-level linkage, and y represents the true value corresponding to each feature vector.

Deep forest is a cascade model between layers. The feature vectors obtained from each layer will be spliced with the original feature vectors as the input of the next layer for layer-by-layer training. The input feature x^{k} received in the k-level cascade is the splicing of the original feature x^{0} and the output x^{k-1} in the $k-1$ cascade, so the combined feature is expressed as,

$$x^{k}=\left(f_{k,1}\left(x^{k-1}\right),\ldots,f_{k,L}\left(x^{k-1}\right),x^{0}\right), \tag{3}$$

where $f_{k,L}(x)$ represents the real value of feature x obtained through the L-th forest in the k-level cascade training.

The final prediction is the average of all forest predictions in the last cascade,

$$f(x)=\frac{\sum_{l=1}^{L} f_{K,l}\left(x^{K-1}\right)}{L}, l=1,2,\ldots,L. \tag{4}$$

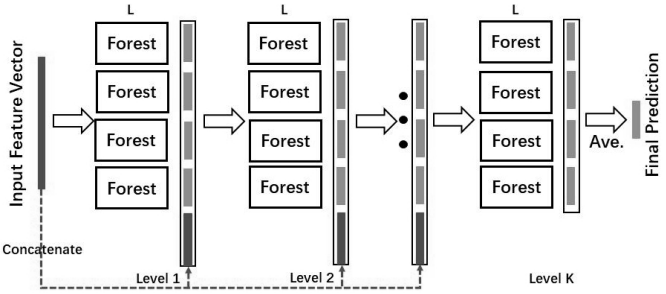

Figure 4. Flow chart of deep forest regression. The predicted value output by the random forest in each hierarchy is spliced with the original feature, which serves as the input of the next hierarchy. The predicted value is output at the last level, and the final predicted value is obtained after the average value is calculated.

The algorithm flow of predicting chemical yield regression prediction based on deep forest is shown in algorithm 2.

Algorithm 2: The intelligent Predicting Reaction Performance in Multi-Dimensional Chemical Space Using Deep regression Forest algorithm.

Input：Training set $S=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, $x_i\in\Re^m$, $y_i\in\{1,\ldots,C\}$.

Where x_i represents the feature descriptor and y_i represents the corresponding yield.

1：**for** $k=1,2,\ldots,K$ **do**

2： **for** $l=1,2,\ldots,L$ **do**

3： According to the deep forest algorithm [35], the L-th random forest of the t-th cascade is trained.

4： Calculate the yield value $f_{k,l}\left(x^k\right)$ of the L-th random forest in the k-th cascade.

5： $x^{k+1}=\left(f_{k,1}\left(x^k\right),\ldots,f_{k,L}\left(x^k\right),x^0\right)$.

6： **end for**

7： Use x^{k+1} for the input of the next cascade.

8：**end for**

Output：Predicted final yield of deep forest $f(x)=\frac{\sum_{l=1}^{L} f_{K,l}\left(x^{K-1}\right)}{L}$.

2.3 Calculation of feature importance

For the classification problem, the importance of features is judged by comparing the contribution of each feature descriptor to each tree in the forest, that is, by comparing the change of Gini coefficient when all features are classified on the base classifier. For the regression problem, the importance of the characteristics is judged by comparing the MSE after division.

Deep forest is an algorithm based on the decision tree. During classification, the nodes of the decision tree will branch and calculate the Gini coefficient of each feature during node division. The smaller the Gini coefficient, the higher the purity of the data set, that is, the better the division effect. The calculation method of Gini coefficient is as follows:

$$Gini(D)=\sum_{k=1}^{n} p_k(1-p_k). \tag{5}$$

where k is the number of classification categories, p_k is the probability that the sample point belongs to class k.

Therefore, the change of Gini coefficient before and after node m branching is the importance of feature X_j on node m:

$$VIM_{jm}^{(Gini)}=GI_m-GI_l-GI_r. \tag{6}$$

where GI_l and GI_r are the Gini indexes of the two new nodes after branching. The smaller the Gini coefficient of the node is, the greater the change of the Gini coefficient before and after division is. Therefore, the greater the difference of the Gini coefficient is, the higher the importance of the current feature is. Calculate the importance of each feature in each tree, and then take the weighted average to obtain the final feature importance evaluation.

In regression prediction, the importance of features is determined by measuring the value of

a given descriptor and the percentage of MSE change after model retraining. The more the percentage of MSE increases, the more important the descriptor is.

3 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. In the classification prediction of yield, accuracy, Error Reduction Rate [24] (ERR) and kappa statistics [25] are selected to measure the classification accuracy of the model. The prediction effect of the model is evaluated from the perspectives of machine learning and statistics.

3.1 Classification evaluation

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.

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{7}$$

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 $\hat{y}_i$.

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 model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{8}$$

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.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{9}$$

3.2 Regression evaluation

Accuracy refers to the proportion of correctly predicted samples in the total samples. The higher the proportion of correctly predicted samples in the total samples is, the higher the classification accuracy is. In the classification problem, the samples can be divided into four cases: true positive (TP), false positive (FP), true negative (TN) and false negative (FN) according to the prediction category of the classifier and the real category of observation. For the T classes classification problem, the total number of samples is.

$$Total = \sum_{t=1}^{T}(TP_t + FP_t). \tag{10}$$

Therefore, the classification accuracy is defined as,

$$Accurary = \frac{\sum_{t=1}^{T} TP_t}{Total}. \tag{11}$$

ERR indicates the percentage of error rate reduction compared with the two methods. The calculation formula for ERR is:

$$ERR = \frac{ER_{Othermethoeds} - ER_{DF}}{ER_{Othermethoeds}} \times 100\%, \tag{12}$$

Where ER is the error rate, that is, 1- accuracy.

Kappa proposed by Cohen [25] can be used not only for consistency test, but also for measuring classification accuracy. The calculation of kappa statistic is based on confusion matrix. P_e represents the sum of "the product of actual samples and predicted samples" corresponding to all categories, divided by "the square of the total number of samples". Kappa statistics is defined as:

$$K = \frac{P_0 - P_e}{1 - P_e} \times 100\%, \tag{13}$$

where P_0 is the sum of the number of correct samples for each category divided by the total number of samples, that is, the overall classification accuracy. The more unbalanced the

confusion matrix, the higher P_e is, the lower the corresponding kappa value is, which can just give a low score to the model with unbalanced category.

The general value range of kappa is (0, 1). When 0 < kappa ≤ 0.2, the consistency is very low; when 0.2 < kappa ≤ 0.4, the consistency is general; when 0.4 < kappa ≤ 0.6, the consistency is medium; when 0.6 < kappa ≤ 0.8, the height is consistent; when 0.8 < kappa < 1, the description is almost identical. The closer the kappa coefficient is to 1, the better the consistency of classification.

4 Results and analysis

By observing the yield distribution of Buchwald-Hartwig coupling reaction (Fig. 5A), most of the yield distribution is 0. According to the fitted probability density curve, the yield distribution is skewed. Furthermore, the Kolmogorov-Smirnov test of the yield showed that the p value was 0.000, which was less than the given significant level α=0.05. Therefore, the yield distribution of Buchwald-Hartwig coupling reaction did not follow the normal distribution. Because it does not satisfy the assumption of normal distribution of general linear model, ML algorithms is more suitable for regression prediction of yield.

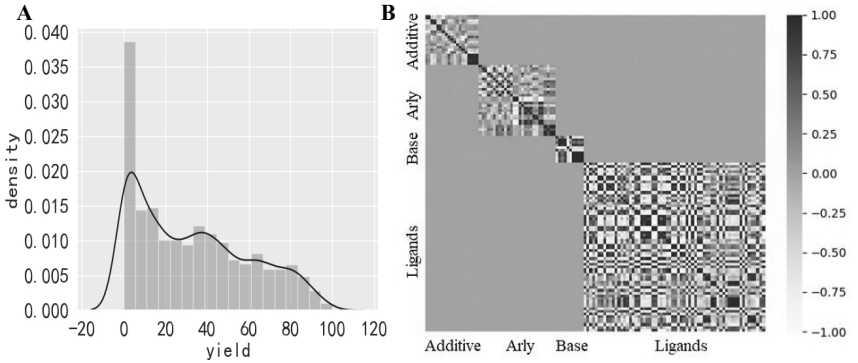

Figure 5. (A)Yield distribution. The frequency distribution histogram and probability density curve of the observed yield values show that the yield is non normal distribution. (B) Correlation plots for descriptors. The relationships among four descriptors, additive, aryl, base, and ligands. There is no correlation among the four descriptors, but a strong correlation within each descriptor.

Figure 5B shows the correlation of four types of descriptors: Additive, Aryl, Base, and Ligands. Each block displays the covariance between the two descriptors. The darker the

color is, the stronger the positive correlation between descriptors is. On the contrary, the lighter the color is, the stronger the negative correlation between descriptors is. Obviously, these four types of descriptors are not related to each other. However, there is a strong correlation within each type of descriptor. Therefore, in order to improve the yield of Buchwald Hartwig coupling reaction, it is necessary to properly match the four types of reactants, including Additive, Aryl, Base, and Ligands.

Due to the strong correlation between each class of descriptors, the traditional machine learning methods usually need to carry out feature screening in processing, which will cause irreversible information loss. The deep forest training feature descriptor is used to predict the yield and yield category. The input feature descriptor is trained layer by layer without feature screening. All information is fully utilized, which can effectively improve the prediction accuracy.

4.1 Yield prediction based on deep forest

In the regression prediction of the yield, all the feature descriptors and the corresponding yield are used as the input of the model. 70% of all the data is selected as the training set, and the remaining 30% is selected as the test set. Deep forest is compared with linear regression (LR) and other seven common ML algorithms, which are k-nearest neighbor [26] (KNN), support vector machine (SVM) [27], neural network (NN) [28], decision tree (DT) [29], random forest (RF) [30], Extra Tree (ET) [31] and convolutional neural network (CNN) [32].

Forecast results are shown in Figure 6. From left to right and from top to bottom are the fitting scatter plots of LR, KNN, SVM, DT, RF, CNN, and DF. The abscissa represents the predicted yield value, and the ordinate represents the observed real yield value. The distribution of each point in the scatter plots of LR, KNN and SVM is far from the fitting line, while almost all the points in the scatter plots of DT, RF and DF are distributed near the fitting line, indicating that the predicted value of the model is very close to the actual value in the test set, the sum of squares of residuals is small, and the distribution of points in DF is more concentrated than that of the other eight machine learning algorithms. It shows that the fitting effect between the yield predicted by deep forest and the real yield is the best among the nine different algorithms, and the regression prediction result of deep forest is better.

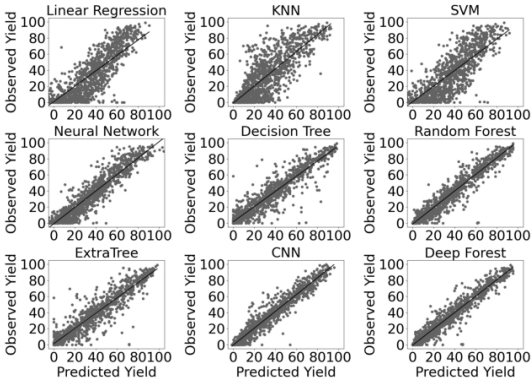

Figure 6. Regression prediction results of test set. For all the models, a 70/30 split of training and test data, with k-fold cross-validation on the training data. Compared the prediction effect of linear regression, ML algorithms, deep learning, and deep forest respectively. The fitting effect of deep forest is the best.

The prediction result of deep forest is that R^2 value is 0.94, RMSE is 6.86 and MAE is 4.52 (Table 1), which is better than ensemble learning and deep learning. This shows that deep forest can accurately predict the yield of Buchwald-Hartwig coupling reaction.

In the classification prediction of yield, the 1 / 4 and 3 / 4 quantiles of yield are defined as the threshold, the yield less than or equal to 1 / 4 quantile is low yield, the yield greater than 1 / 4 quantile and less than or equal to 3 / 4 quantile is medium yield, and the yield greater than 3 / 4 quantile is high yield. After the output of the three categories is balanced by SMOTE [33], all feature descriptors and corresponding categories are used as the input of the model. For all the models, there is a 70/30 split of training and test data.

Table 1. Comparison of regression prediction results.

	R^2	RMSE	MAE
LR	0.69	15.3	12.43
KNN	0.62	17.0	12.94
SVM	0.68	15.8	12.48
NN	0.88	9.6	7.24
DT	0.88	9.7	6.00
RF	0.92	7.7	5.07
ET	0.90	8.7	6.16
CNN	0.93	7.3	5.19
DF	**0.94**	**6.8**	**4.52**

Similarly, deep forest is compared with logistic regression (LR) and other six common ML algorithms, which are KNN, SVM, NN and RF. Table 2 measures the classification effect of each algorithm by comparing the accuracy, error reduction rate (ERR) and kappa statistics. Among them, the classification accuracy of deep forest is the highest, which is 93.72%. Compared with other algorithms, the error rate of DF is reduced in varying degrees. In addition, kappa consistency test shows that the kappa value of deep forest is the largest and greater than 0.8, which also proves that the predicted classification results are basically consistent with the actual classification results from a statistical point of view. In conclusion, it shows that deep forest can accurately predict the yield of Buchwald Hartwig coupling reaction.

Table 2. Comparison of classification prediction results.

	accuracy	ERR	kappa
LR	70.82%	↓78.48%	0.562
KNN	77.78%	↓71.74%	0.666
SVM	55.56%	↓85.87%	0.329
NN	63.58%	↓82.76%	0.447
DT	86.81%	↓52.39%	0.802
RF	91.58%	↓25.42%	0.874
ET	83.33%	↓62.33%	0.750
DF	**93.72%**	—	**0.906**

Table 3. Significance test of classification accuracy of different algorithms.

	Friedman
Chi-square statistic	37.071
p-value	0.0000

In order to test whether there is significant difference in classification prediction between the total algorithms, calculate the classification accuracy of different algorithms when the test set accounts for 30%, 40%, 50%, 60%, 70%, 80% and 90% of all data (the results are shown in table S1). Friedman test is carried out on them, and the chi square measurement is 37.071 and the p value is 0.000 (Table 3). Therefore, the original hypothesis is rejected. It shows that

there are significant differences in classification accuracy between different algorithms.

Comparing the classification results of deep forest and convolutional neural network, it can be found that the classification accuracy of deep forest is higher and the prediction time is shorter (Table 4). The experimental results show that compared with deep learning, deep forest develops its strengths and circumvents its weaknesses. It has good performance in small sample prediction and runs fast under CPU.

Table 4. Comparison between DF and CNN.

	DF	CNN
accuracy	93.72%	89.45%
err	—	↓40.47%
kappa	0.906	0.842
time	8s	118.81s

To sum up, deep forest is also very competitive compared with convolutional neural network. The former also has good performance when the amount of data is small, and has fewer super parameters. Generally, better prediction results can be obtained by using the default super parameters, and it is easier to optimize parameters in various tasks. In addition, the running time of deep forest is shorter, which can effectively save time in prediction.

4.2 Analysis of influencing factors of yield

In order to further understand the effect of different feature descriptors on the yield of Buchwald-Hartwig, deep forest was used to calculate the feature importance of each descriptor and rank it. By measuring the characteristic importance of the descriptors, the main factors affecting the yield were analyzed, and the relationship between the reaction conditions of Buchwald-Hartwig and yield was searched.

According to the ranking of characteristic importance (Fig. 7A), aryl halide accounted for 5. It is the most important factor affecting the regression prediction of yield, including aryl halide *C-3 nuclear magnetic resonance (NMR) shift, aryl halide *H2 electrostatic charge, aryl halide vibration descriptor (V2, V3) and molecular descriptor. In addition, additive

accounted for 3 of the top 10 descriptors. It is the second important factor that affects the regression prediction of yield, including additive *C-3 NMR shift and *O-1 and *C-5 electrostatic charges.

In the same way, according to the ranking of characteristic importance (Fig. 7B), aryl halide and additive descriptors are the most important factors affecting the yield classification prediction. The composition classification of aryl halide is the same as that of regression. Additives mainly include additive electronic descriptors (*C-4, *C-3, *O-1) and additive *C-3 NMR shift, *C-4 NMR shift. Therefore, the aryl halide and additive electronic properties appear to be important in yield prediction of Buchwald Hartwig coupling reaction. Additive electronic charge describes the electron rich degree. In general, the higher the electron rich degree of additive is, the lower the yield is.

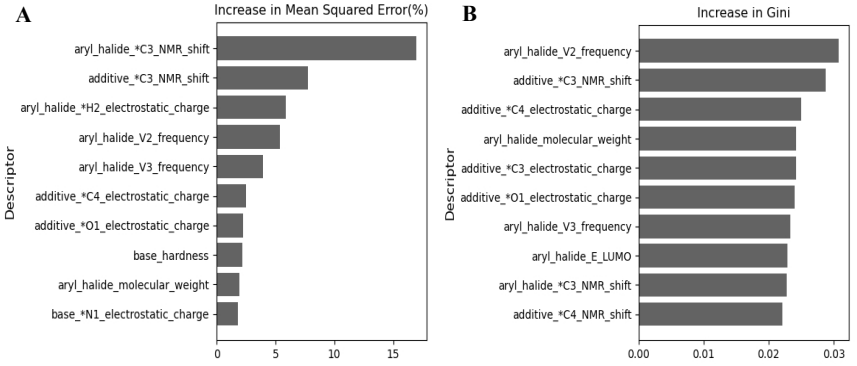

Figure 7. The importance of feature descriptors. (A) The 10 most important descriptors of deep forest in regression prediction, and it is determined by comparing the increase percentage of MSE value when the given descriptor value is recombined and the model is retrained. (B) The 10 most important descriptors of deep forest in classification prediction, and it is determined by comparing the increase of Gini index during training. * indicates a shared atom. E. energy; Homo, highest occupied molecular orbital; V. vibration.

4.3 Analysis of influencing factors of yield

Finally, the 23 kinds of isoxazole additives are divided into training set (1st to 17th and 20th, 23rd) and test set (18th, 19th, 21st and 22nd) to make out-of-sample prediction for additive, and isoxazole in training set is used to predict the performance of isoxazole in test set (Fig. 8). The purpose is to explore the effect of a new isoxazole additive on the yield of

Buchwald-Hartwig coupling reaction.

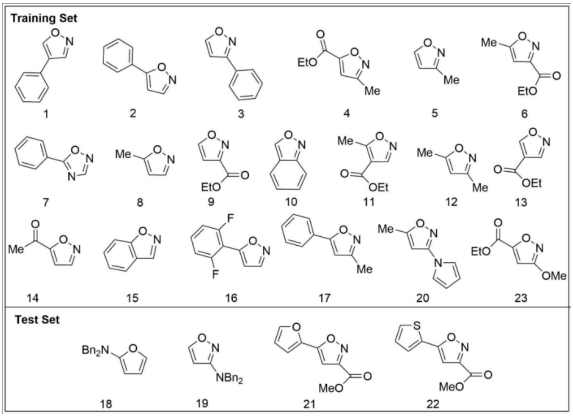

Figure 8. Structure of additive performance prediction. Using isoxazole in training set (1st to 17th and 20th, 23rd) to predict the performance of isoxazole in test set (18th, 19th, 21st and 22nd).

In the regression prediction, the out-of-sample prediction results of the additive are shown in Figure 9. Almost all the points in the scatter diagram of the four additives (18th, 19th, 21st and 22nd) are distributed near the fitting line, which indicates that the predicted yield of the model is very close to the actual yield in the test set, and RMSE of the residuals is small.

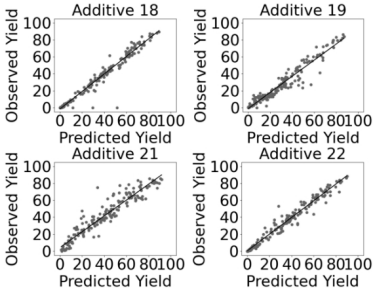

Figure 9. Performance prediction of additive. The fitting scatter plots of predicted values and true values of the 18th, 19th, 21st and 22nd isoxazole additive test sets were compared respectively. The closer the distribution of scatter points to the fitting line, the better the fitting effect.

Similarly, the out of sample prediction of additives also shows high accuracy (Table 5). All results show that there is no significant deviation between the predicted value outside the sample and the observed value of the additive. Deep forest can effectively predict the Buchwald-Hartwig coupling reaction and can predict the influence of reactants with unknown

properties on the reaction, to provide the highest yield.

Table 5. Additive classification prediction.

	accuracy	kappa	micro_f1
Additive18	93.89%	0.905	0.939
Additive19	88.89%	0.802	0.846
Additive21	93.89%	0.899	0.933
Additive22	94.44%	0.914	0.939

5 Conclusions

Deep forest combines the advantages of deep learning and ensemble learning, and can be used for regression and classification prediction with higher accuracy than traditional ML algorithms. An intelligent prediction and analysis method for Buchwald-Hartwig coupling reaction based on deep forest is proposed in this paper, which can realize high precision prediction of regression and classification of yield. By calculating the importance of characteristic descriptors, the factors significantly affecting the yield prediction of Buchwald-Hartwig coupling reaction can help chemists to make decisions, improve the yield, and promote the rapid realization of organic synthesis. The results of out of sample prediction can be used to predict the properties of reactants with unknown properties (such as additives), to help chemists quickly understand the reaction properties of unknown compounds and how to affect the yield of Buchwald Hartwig coupling reaction. The synthesis of complex molecules usually takes a lot of time and resources. This method conforms to the concept of green chemistry and improves the yield of Buchwald-Hartwig coupling reaction under the condition of reducing the cost.

References

[1] E. Fischer, F. Jourdan, Ueber die Hydrazine der Brenztraubensäure, *Ber. Dtsch Chem. Ges*. **16** (1883) 2241–2245.

[2] F. Ullmann, Ueber eine neue Bildungsweise von Diphenylaminderivaten, *Ber. Dtsch. Chem. Ges*. **36** (1903) 2382–2384.

[3] I. Goldberg, Ueber Phenylirungen bei Gegenwart von Kupfer als Katalysator, *Ber. Dtsch. Chem. Ges*. **39** (1906) 1691–1692.

[4] F. Paul, J. Patt, J. F. Hartwig, Palladium-catalyzed formation of carbon- nitrogen bonds. Reaction intermediates and catalyst improvements in the hetero cross-coupling of aryl halides and tin amides, *J. Am. Chem. Soc.* **116** (1994) 5969–5970.

[5] A. S. Guram, S. L. Buchwald, Palladium-catalyzed aromatic aminations with in situ generated aminostannanes, *J. Am. Chem. Soc.* **116** (1994) 7901–7902.

[6] P. Ruizcastillo , D. G. Blackmond, S. L. Buchwald, Rational ligand design for the arylation of hindered primary amines guided by reaction progress kinetic analysis, *J. Am. Chem. Soc.* **137** (2015) 3085–3092.

[7] A. B. Lopes, P. Wagner, Rodrigo Octavio Mendonça Alves de Souza, N. L. Germain, J. Uziel, J. Bourguignon, M. Schmitt, L. S. M. Miranda, Functionalization of 2H-1,2,3-triazole C-nucleoside template via N2 selective arylation, *J. Org. Chem*. **81** (2016) 4540–4549.

[8] X. Y. Zhao, Q. Zhou, J. M. Lu, Synthesis and characterization of N-heterocyclic carbene-palladium (II) chlorides-1-methylindazole and -1-methylpyrazole complexes and their catalytic activity toward C–N coupling of aryl chlorides, *RSC Advances* **6** (2016) 24484–24490.

[9] C. C. C. Johansson Seechurn, T. Sperger, T. G. Scrase, F. Schoenebec, T. J. Colaco, Understanding the unusual reduction mechanism of Pd(II) to Pd(I): Uncovering hidden species and implications in catalytic cross-coupling reactions, *J. Am. Chem. Soc*. **139** (2017) 5194–5200.

[10] J. Clark, C. Voth, M. J. Ferguson, M. Stradiotto, Evaluating 1,1′-bis (phosphino) ferrocene ancillary ligand variants in the nickel-catalyzed C–N cross-coupling of (hetero) aryl chlorides, *Organometallics* **36** (2017) 679−686.

[11] A. Yada, K. Nagata, Y. Ando, T. Matsumura, S. Ichinoseki, K. Sato, Machine learning approach for prediction of reaction yield with simulated catalyst parameters, *Chem. Lett*. **47** (2018) 284–287.

[12] M. Fujinami, J. Seino, H. Nakai, Quantum chemical reaction prediction method based on machine learning, *Bull. Chem. Soc. Jpn*. **93** (2020) 685–693.

[13] R. Burbidge, M. Trotter, B. Buxton, S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, *Comput. Chem*. **26** (2001) 5–14.

[14] M. H. Fatemi, A. Heidari, M. Ghorbanzade, Prediction of aqueous solubility of drug-like compounds by using an artificial neural network and least-squares support vector machine, *Bull. Chem. Soc. Jpn*. **83** (2010) 1338–1345.

[15] R. Zhang, X. Li, X. Zhang, H. Qin, W. Xiao, Machine learning approaches for elucidating the biological effects of natural products, *Nat. Product Rep*. **38** (2021) 346–361.

[16] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, S. Venkatesh, GraphDTA: Predicting drug–target binding affinity with graph neural networks, *Bioinf*. **37** (2021) 1140–1147.

[17] K. Hatakeyama-Sato, T. Tezuka, Y. Nishikitani, H. Nishide, K. Oyaizu, Synthesis of lithium-ion conducting polymers designed by machine learning-based prediction and screening, *Chem. Lett*. **48** (2019) 130–132.

[18] M. Fujinami, J. Seino, T. Nukazawa, S. Ishida, T. Iwamoto, Virtual reaction condition optimization based on machine learning for a small number of experiments in high-dimensional continuous and discrete variables, *Chem. Lett*. **48** (2019) 961–964.

[19] E. Shimono, T. Kurita, Y. Ichiraku, Logistic regression analysis for the material design of chiral crystals, *Chem. Lett*. **47** (2018) 611–612.

[20] S. Shojiro, F. Kimito, Industrial case study: Identification of important substructures and exploration of monomers for the rapid design of novel network polymers with distributed representation, *Chem. Lett*. **94** (2021) 112–121.

[21] 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.

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

[23] Z. Zhou, J. Feng, Deep forest, *Nat. Sci. Rev*. **6** (2019) 74–86.

[24] W. Deng, J. Hu, J. Guo, Extended SRC: under sampled face recognition via intraclass variant dictionary, *IEEE Trans. Patt. Anal. Mach. Intell*. **34** (2012) 1864–1870.

[25] J. Cohen, A coefficient of agreement for nominal scales, *Educ Psychol Meas*. **20** (1960) 37–46.

[26] M. Aci, M. Avci, K-nearest neighbor reinforced expectation maximization method, *Expert Syst. Appl*. **38** (2011) 12585–12591.

[27] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, D. Haussler, Support vector machines classification and validation of cancer tissue samples using microarray expression data, *Bioinf*. **16** (2000) 906–914.

[28] S. S. Haykin, *Neural Networks – A Comprehensive Foundation*, Pearson, Delhi, 1994.

[29] J. Quinlan, Induction of decision trees, *Mach. Learn*. **1** (1996) 81–106.

[30] L. Breiman, Random forests, *Mach. Learn*. **45** (2001) 5–32.

[31] P. Geurts, Extremely randomized trees, *Mach Learn*. **63** (2006) 3–42.

[32] Y. Lecun, Y. Bengio, *Convolutional Networks for Images, Speech, and Time-Series*, MIT Press, Cambridge, 1995.

[33] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, *J. Artif. Intell. Res.* **16** (2002) 321–357.