MATCH
Communications in Mathematical
and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **87** (2022) 299-318
ISSN: 0340-6253
doi: 10.46793/match.87-2.299P

Intelligent Predicting Reaction Performance in Multi-Dimensional Chemical Space Using Quantile Regression Forest

Lichao Peng[1], Jing Dong[2], Xuechun Mu[2], Zelin Zhang[3], Yuqing Zhang[4], Xiaohui Yang[*2], Puyu Zhang[*4]

[1]National & Local Joint Engineering Research Center for Applied Technology of Hybrid Nanomaterials, Henan University, Kaifeng, China, 475000
[2]Henan Engineering Research Center for Artificial Intelligence Theory & Algorithms, School of Mathematics and Statistics, Henan University, Kaifeng, China, 475000
[3]School of Computer and Information Engineering, Henan University, Kaifeng, China, 475000
[4]College of Chemistry and Chemical Engineering, Henan University, Kaifeng, China, 475000
xhyanghenu@163.com, zhangpuyu@henu.edu.cn

(Received August 17, 2021)

Abstract

Buchwald-Hartwig amination reaction is widely applied in synthetic organic chemistry, which faces tedious and complex experimental process. In 2018, an interesting yield prediction technique is proposed via machine learning (random forest) in ***Science***. However, the method is based on point prediction with many feature descriptors. For tackling these problems, complements and improvements have been made from the perspectives of machine learning and statistics, including feature dimensionality reduction, distributed prediction and visualization, so as to provide accurate and reliable decision information.

1 Introduction

Chemical reaction research is often faced tedious experimental process and extensive data research. Machine learning (ML) can not only solve these problems well, but also find the relationship between data in a large amount of chemical information and help researchers to make reasonable judgments and decisions. ML has been increasingly favored by many chemists and has made progresses in computer synthetic design system [1,2], such as chemical reaction performance prediction [3-8], drug research and development [9,10], auxiliary high-performance materials design [11], auxiliary inverse synthesis analysis [12-14], and screening target compound [15].

Coupling reaction is very important in organic synthesis, and its products are widely used in medicine, pesticide, natural products, and even advanced functional materials. The palladium-catalyzed cross coupling reaction is a kind of coupling reaction, which refers to the reaction with palladium compound as catalyst (mostly homogeneous catalyst). As early as 1972, Richard F. Heck found that palladium could be used as a catalyst to realize the connection between carbon atoms under mild conditions [16]. Ei-ichi Negishi and Akira Suzuki further developed the methods of cross-coupling C-C atoms catalyzed by palladium in 1977 [17,18] and 1979 [19,20], respectively, so that the substrates and product types of such chemical reactions were further expanded. These methods make it simple and efficient to make stable carbon atoms easily join together to form more complex molecules. With the advent of these cross-coupling methods, chemists were able to manipulate atoms and molecules at an unprecedented level.

Palladium catalysis can not only realize the coupling reaction of C-C binding, but also realize the coupling reaction of carbon-heteroatom binding. The formation of C-N is an important field in modern organic synthesis. Amines and their derivatives, nitrogen-containing heterocycles, etc, can be prepared by forming C-N bonds. Many of them are compounds with biological and medicinal activities and some important intermediates. Buchwald-Hartwig amination reaction is an efficient and universal method for the synthesis of substituted aromatic amines and is also one of the research hot points in the field of organic synthesis of C-N bond catalyzed by palladium [21-24].

In recent years, the development of ML algorithms provides a "shortcut" for Buchwald-Hartwig amination reaction to find the appropriate substrate and reaction conditions. This method has become an integral part of scientific inquiry in many disciplines and has also brought new opportunities for the development of organic chemistry. By screening or coding the information in the chemical system to form a certain expression of chemical information, that is descriptor, the research in the chemical field can be transformed into a process of data processing, so as to reduce the dependence on personnel in a certain program. ML methods can mine the correlation of massive experimental data generated in chemical experiments, help chemists make reasonable analysis and prediction, and greatly accelerate the chemical research and development process.

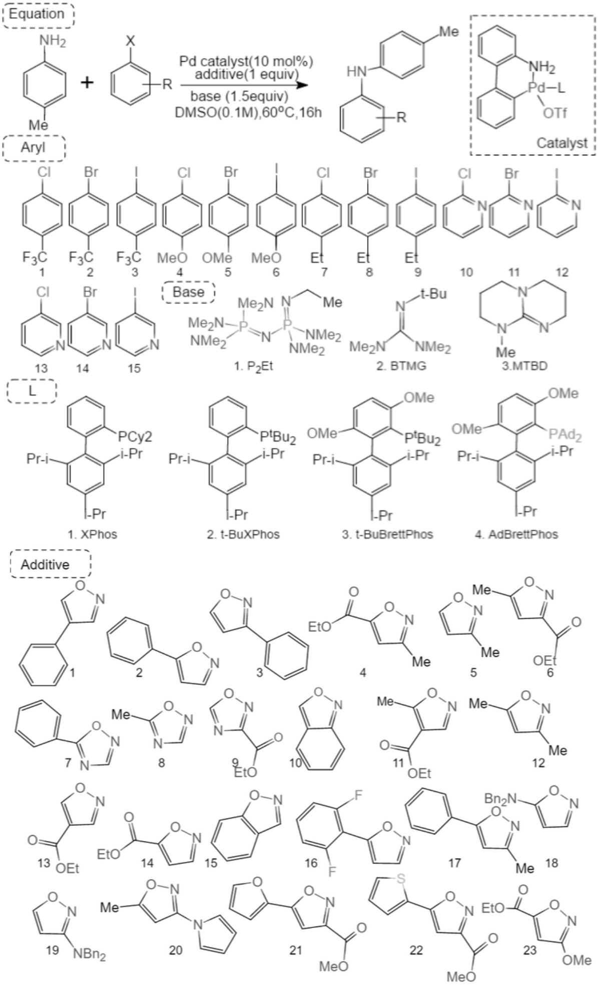


Figure 1. The reaction components of Buchwald-Hartwig amination reaction. Here Me is methyl, X is any halide, L is ligands, OTf is triflate, t-Bu is tert butyl, and i-Pr is isopropyl.

Doyle et al reported the prediction of Buchwald-Hartwig amination yield by random forest model [3]. They selected the Pd-catalyzed Buchwald-Hartwig reaction as test reaction for model development because of its broad value in pharmaceutical synthesis. Nevertheless, the application of this reaction to complex drug-like molecules remains challenging [25]. One limitation is the poor performance of substrates possessing five-membered heterocycles, such as isoxazoles. These heterocycles have drug-like characteristics but are underrepresented in successful drug candidates [26]. Thus, the author sought to use ML to predict the performance of the Buchwald-Hartwig reaction in the presence of isoxazoles. The database are 4608 reaction data consisting of 23 isoxazole additives, 15 aryl and heteroaromatic halides, ligand of 4 palladium catalysts, and 3 base substrates. Figure 1 gives the reaction structure diagram similar to Doyle et al [3]. The author used 120 descriptors (atomic descriptors (64), molecular descriptors (28) and vibration descriptors (28)) as input and reaction yield as output to predict the yield using random forest model, and the predicted results were R2=0.92, RMSE=7.80. The interesting research provides a powerful tool for chemists to guide chemical synthesis methods and predict chemical reaction properties.

However, the prediction of chemical reaction performance is limited to point estimation-based point prediction, which only depicts the average level of prediction results. Moreover, point estimation is an inference method that does not consider sampling error and directly replaces all indexes with sampling index. It is inevitable that there will be errors in directly replacing all indicators with sampling indicators. Apart from point estimation, interval estimation is also used in parameter estimation. Interval estimation is an inference method to estimate the possible range of all indicators according to sampling index and sampling error in sampling inference. In inferring all indicators from the sampling index, a certain probability is used to ensure that the error does not exceed a given range [27]. The range of the estimated value can be judged at a certain probability level, so as to understand the degree of aggregation and discrete of the sample sequence.

Based on the above, we complements and improvements from another perspective that article by Doyle et al, which are mainly divided into the following aspects:

(1) the performance of ML model depends on the effective feature representation corresponding to the data itself, that is, a comprehensive and concise feature descriptor. Based on the 120 feature descriptors in Doyle [3], we propose a method to optimize feature descriptors to describe information with as few descriptors as possible.

(2) we perform an interval prediction and a probability density curve fitting analysis of the yield prediction of the chemical reaction from a statistical perspective.

Compared with Doyle [3], this paper aims to give a more concise and effective description of intelligent prediction of chemical synthesis reaction from the perspective of machine learning and statistics. The flowchart is shown in Figure 2.

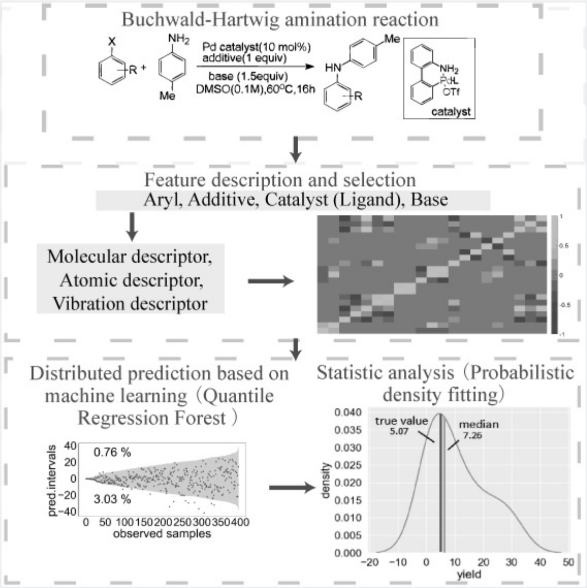

Buchwald-Hartwig amination reaction
Pd catalyst(10 mol%)
additive(1 equiv)
base (1.5equiv)
DMSO(0.1M),60°C,16h
catalyst
Feature description and selection
Aryl, Additive, Catalyst (Ligand), Base
Molecular descriptor,
Atomic descriptor,
Vibration descriptor
Distributed prediction based on machine learning（Quantile Regression Forest）
Statistic analysis（Probabilistic density fitting）
0.76 %
3.03 %
pred.intervals
observed samples
true value 5.07
median 7.26
density
yield

Figure 2. Flowchart of intelligent prediction of chemical synthesis reaction.

2 Integrated feature selection based on importance and relevance

Due to the high dimension of data itself, data should be screened before model training, and useful data should be selected as the input of the model. However, in the data selection, there is no strict conclusion about how much to choose and how to choose. Therefore, it is the pursuit of the goal to use comprehensive and concise characteristic data to achieve the highest prediction accuracy. Therefore, we propose an integrated feature selection method based on importance and relevance for descriptor data of chemical reactions.

In order to obtain more comprehensive feature information, we selected Between category to within-category sums of squares (BW), used to "score" the features through a certain index, and then the features were sorted according to the scores, and finally the Top K features were selected as information features [28]. BW for a feature j, the score is defined as:

$$BW(j)=\frac{BSS(j)}{WSS(j)}=\frac{\sum_{i}\sum_{k}I(l_i=k)(\bar{x}_{kj}-\bar{x}_{.j})^2}{\sum_{i}\sum_{k}I(l_i=k)(\bar{x}_{ij}-\bar{x}_{kj})^2}, \tag{1}$$

where $\bar{x}_{.j}$ is the average expression level of the j-th feature of all samples; $\bar{x}_{kj}$ is the average expression level of the j-th feature of all samples of the k-th class; l_i is the class label of the current sample; $I(*)$ is a discriminant function used to determine which category the current corresponding sample belongs to. When the * logical expression is true, its value is 1, otherwise, it is 0.

After the preliminary feature screening, there may still be a strong correlation between the features. Therefore, in order to obtain as few chemical descriptors as possible and achieve the highest prediction accuracy, we used Least Absolute Shrinkage and Selection Operator (LASSO) to screen the features obtained from the preliminary screening again to remove the correlation [29].

LASSO is a regression analysis method proposed by Tibshirani in 1996 [29], which can not only select variables, but also achieve regularization. The basic idea is to minimize the sum of squares of residuals under the constraint that the l_1 norm of the regression coefficient is less than a constant, so that some regression coefficients which are strictly equal to 0 can be generated. Let the data $\{(x_i, y_i); i=1,2,..., N\}, x_i \in R^d$, where x_i, y_i are respectively the regression quantity of the *i*-th observed value and the corresponding label, $\beta=(\beta_1,\beta_2,...,\beta_d)^T$ is the regression coefficient vector, and β_0 is the intercept. The objective function of Lasso is:

$$\begin{cases} (\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0, \beta} \sum_{i=1}^{N} (y_i - x_i^T \beta - \beta_0)^2 \\ \quad\quad s.t. \ \sum_{i=1}^{d} \left|\beta_j\right| \le \lambda, \end{cases} \tag{2}$$

where λ is the penalty parameter, which is the weight of the size of the coefficient and controls the degree of shrinkage. The larger λ is, the greater the degree of contraction is, and vice versa.

3 Quantile regression forest and kernel density estimation

Nicolai Meinshausen proposed an interval estimation method named Quantile Regression Forest (QRF) [30], which preserves all observations of the node, not just the average, and can assess the distribution of conditions based on this information. According to the prediction interval, outliers can also be obtained. Probability density prediction can provide the complete probability density curve of the sample and provide more effective information to the researchers. Therefore, we select this model to predict the yield of chemical reaction.

3.1 Quantile regression forest

The core idea of quantile regression is to generalize from the mean to the quantile. The goal of least square regression is to minimize the mean square error (MSE), while the goal of quantile regression is equivalent to a weighted least square method [31]. The following type:

$$\begin{aligned} & MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = E(y_i - \hat{y}_i)^2 \\ & \hat{Q}_y(\tau) = \arg\min(\sum_{i:y_i \ge \hat{y}_i^\tau} \tau\left|y_i - \hat{y}_i^\tau\right| + \sum_{i:y_i < \hat{y}_i^\tau} (1-\tau)\left|y_i - \hat{y}_i^\tau\right|) \end{aligned} \tag{3}$$

where τ is the quantile selected; y_i is the actual value of the sample i; $\hat{y}_i$ is the predicted value; $\hat{y}_i^\tau$ is the predicted value of the i samples in the τ quantile.

Nicolai Meinshausen proposed quantile regression forest algorithm in 2006 using the principle of quantile regression, the consistency of the algorithm is proved mathematically. One application of QRF is to construct prediction intervals $I(x)$ [30]. Let Y be a real-valued response variable and X a covariate or predictor variable, possibly high-dimensional. The conditional distribution function $\mathrm{F}(y|X = x)$ is given by the probability that, for $X = x$, Y is smaller than $y \in R$, $\mathrm{F}(y|X = x) = P(Y \le y|X = x)$. For a continuous distribution function, the α -quantile $Q_\alpha(x)$ is then defined such that the probability of Y being smaller than $Q_\alpha(x)$ is, for a given X=x, exactly equal to α. For each new data point X, a prediction interval of :

$$\begin{aligned} &I(x)=[Q_{.05}(x),Q_{0.95}(x)],\\ &Q_\alpha(x)=\inf\{y:F(y|X=x)\geq\alpha\}, \end{aligned} \tag{4}$$

gives a range that will cover the new observation of the response variable Y with high probability [30]. The steps of the algorithm are as follows:

a) Grow k trees $T(\theta_t)$, $t=1,\cdots,k$, as in random forests. However, for every leaf of every tree, take note of all observations in this leaf, not just their average.

b) For a given $X=x$, drop x down all trees. Compute the weight $w_i(x,\theta_t)$ of observation $i\in\{1,\ldots,n\}$ for every tree as in:

$$w_i(x,\theta_t)=\frac{1_{\{X_i\in R_{l(x,\theta)}\}}}{\#\{j:X_j\in R_{l(x,\theta)}\}}. \tag{5}$$

Compute weight $w_i(x)$ for every observation $i\in\{1,\ldots,n\}$ as an average over $w_i(x)$, $i\in\{1,\ldots,k\}$, as in:

$$w_i(x)=\frac{1}{k}\sum_{t=1}^{k}w_i(x,\theta_t). \tag{6}$$

c) Compute the estimate of the distribution function as in: $\hat{F}(y|X=x)=\sum_{i=1}^{n}w_i(x)1_{\{Y_i\leq y\}}$ for all $y\in R$, using the weights from Step b).

3.2 Kernel density estimation

Kernel Density Estimation (KDE), also known as Parzen window [32]. It is a very effective nonparametric estimator for the unknown density function. Suppose $x_i(i=1,2,\ldots,N)$ is the sample of random variable x, and $f(x)$ is the probability density function of random variable x, then the kernel density estimation at the given point x is expressed as:

$$\hat{f}(x)=\frac{1}{Nh}\sum_{i=1}^{N}K\left(\frac{x-x_i}{h}\right) \tag{7}$$

among them, the h is bandwidth, the n is sample size and the $K(\cdot)$ is kernel function.

Common kernel functions are: Uniform kernel, Epanechnikov kernel, Gaussian kernel and so on. Epanechnikov kernel is optimal in the sense of mean square error, and the efficiency loss is also small. Therefore, the kernel function selected in this paper is Epanechnikov kernel function, and its expression is:

$$K(u)=\frac{3}{4}\left(1-u^{2}\right)\cdot I(|u|\leq 1) \tag{8}$$

where $I(\cdot)$ is the indicative function.

Compared with the kernel function, the influence of bandwidth on the probability density function is greater. When the bandwidth h is small, the curve of kernel density estimation is not smooth and shows the multi-peak feature which the original probability density function does not have; When the bandwidth h is large, the KDE curve is smooth, but it is easy to hide the details. The average integral square error is used to measure the advantages and disadvantages of h , and the optimal bandwidth can be obtained when the minimum is. The formula is as follows:

$$MISE(h)=E\int\left(\hat{f}(x)-f(x)\right)^{2}dx \tag{9}$$

For convenience, the integrated Intelligent Predicting Reaction Performance in Multi-Dimensional Chemical Space Using Quantile Regression Forest, the corresponding algorithm is given as follows and shown in Algorithm 1.

Algorithm 1: The Intelligent Predicting Reaction Performance in Multi-Dimensional Chemical Space Using Quantile Regression Forest algorithm.

```
Input: Feature-descriptor data： $X = [X_1, X_2, \cdots, X_m]^T = [X_M, X_A, X_V]$,  where m represents the number of samples,
    and N represents the number of feature descriptors;
    $X_M$ ： Indicates the molecular descriptor data;
    $X_A$： Indicates the atom descriptor data;
    $X_V$： Indicates the vibration descriptor data.
    $X_i = [x_{i1}, x_{i2}, \cdots, x_{im}], (i = 1,2,\ldots, n)$.
    Yield: $Y=[y_1, y_2, \cdots, y_m]^T$.
Integrated feature selection based on importance and relevance
  1) By calculating the BW score for the $X_M, X_A, X_V$, the appropriate score threshold is selected for the filter
     descriptors, and obtain $X_M^{new}, X_A^{new}, X_V^{new}$.
  2) Further perform the LASSO algorithm based on the filtered data from 1) and the Y, to further remove the
     inter-data correlation
       $\begin{cases} (\hat{\beta}_0, \hat{\beta}) = \arg\min_{\beta_0,\beta} \sum_{i=1}^{N} (y_i - x_i^T \beta - \beta_0)^2 \\ s.t. \ \sum_{i=1}^{d} |\beta_j| \leq \lambda, \end{cases}$
  then, obtain the $X^{new}$.
Quantile Regression Forest and Kernel Density Estimation
  1) Enter the $X^{new}, Y$ into the QRF model:
     a) Grow $k$ trees $T(\theta_t)$, $t = 1, \cdots, k$ , as in random forests. However, for every leaf of every tree, take note
        of all observations in this leaf, not just their average.
```

b) For a given $\hat{X}^{new} = x$, drop x down all trees. Compute the weight $w_i(x,\theta_t)$ of observation $i \in \{1,\ldots,n\}$ for every tree as in:

$$w_i(x,\theta_t) = \frac{1_{\{X_i \in R_{l(x,\theta)}\}}}{\#\{j : X_j \in R_{l(x,\theta)}\}}.$$

Compute weight $w_i(x)$ for every observation $i \in \{1,\ldots,n\}$ as an average over $w_i(x)$, $i \in \{1,\ldots,k\}$, as in:

$$w_i(x) = \frac{1}{k}\sum_{t=1}^{k} w_i(x,\theta_t).$$

c) Compute the estimate of the distribution function as in: $\hat{F}(y|\hat{X}^{new} = x) = \sum_{i=1}^{n} w_i(x)1_{\{Y_i \le y\}}$ for all $y \in R$, using the weights from Step b).

2) For any sample, the KDE function at its different quantile points is calculated:

$$\hat{f}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right),$$

where the h is bandwidth, the $K(\cdot)$ is kernel function: $K(u) = \frac{3}{4}(1 - u^2) \cdot I(|u| \le 1)$.

Output: Different interval estimation results and the probability density curves for any sample.

4 Evaluating indicators

Point prediction indicators are generally used Mean Absolute Percent Error and Root Mean Square Error. The probabilistic prediction indicators are generally reliability indicators (prediction interval coverage) and clarity indicators (average width of forecast interval). Quantile Score (QS) [33] and Winkler Score (WS) were used to evaluate the probabilistic prediction effect of the model [34] in this paper. The QS synthetically considers the reliability index and the definition index, effectively solves the contradiction problem of high confidence level and narrow interval width, the smaller the QS, the better the probability prediction effect is. WS is the interval prediction index, the smaller the value, the higher the interval prediction accuracy. Through these two indicators, the predicted value of the original data obtained by the random forest was compared with the predicted value obtained by QRF based on the 21 descriptors obtained by re-screening, which verified the accuracy of the probabilistic prediction model proposed in this paper. The evaluation indicators appearing in this paper are explained as follows.

(1) The R^2 also called coefficient of determination, is also called the optimal degree of fit, and reflects the degree that the independent variable x explains the changes in the dependent variable y. The closer to 1, the better the model fits.

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2} = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}, \tag{10}$$

where TSS (Total Sum of Squares) $= \sum_{i=0}^{n}(y_i - \bar{y})^2$, it represents the degree of y change, proportional to the variance. RSS (Residual Sum of Squares) $= \sum_{i=0}^{n}(y_i - \hat{y})^2$, it represents the residue of the model and the real values. ESS (Explained Sum of Squares) $= \sum_{i=0}^{n}(\hat{y}_i - \bar{y})^2$, it represents the prediction of the model for the change of the y.

(2) The RMSE (Root Mean Square Error) corresponding to the expectation of the square (quadratic) error. RMSE can accurately calculate the error size of the predicted results and the real results, and can guide our model improvement work such as tuning, feature selection, etc.

$$RMSE(y,\hat{y}_i) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|y_i - \hat{y}_i\|_2^2} \tag{11}$$

(3) The Quantile Score (QS) or pinball loss [33] considers reliability index and clarity index to evaluate the effect of probabilistic prediction, which can solve the contradiction between high confidence level and narrow interval width. The smaller the quantile fraction is, the better the effect of probabilistic prediction is. Such as:

$$QS = \frac{1}{N \cdot card(Q)} \sum_{\tau=Q} \sum_{i=1}^{N} \rho_\tau (y_i - \hat{y}_i^\tau) \tag{12}$$

where, Q is the set of quantiles; card(Q) is the number of quantiles; y_i is the actual value of the i sample; $\hat{y}_i^\tau$ is the predicted value of the i sample of the τ quantile.

(4) Winkler Score (WS) [34] is an interval prediction index, the smaller the value, the higher the interval prediction accuracy. For example:

$$WS = \frac{1}{N}\sum_{i=1}^{N}\begin{cases} (u_i^{1-\alpha} - l_i^{1-\alpha}), & y_i \in \left[l_i^{1-\alpha}, u_i^{1-\alpha}\right] \\ (u_i^{1-\alpha} - l_i^{1-\alpha}) + 2(l_i^{1-\alpha} - y_i)/\alpha & y_i < l_i^{1-\alpha} \\ (u_i^{1-\alpha} - l_i^{1-\alpha}) + 2(y_i - u_i^{1-\alpha})/\alpha & y_i > u_i^{1-\alpha} \end{cases} \tag{13}$$

where $u_i^{1-\alpha}, l_i^{1-\alpha}$ are the upper bound and lower bound of the i sample prediction interval at the confidence level $1-\alpha$ respectively; y_i is the actual value of the i sample; α is the significance level.

5 Results and analysis

5.1 Performance analysis of feature screen

To obtain comprehensive and concise feature descriptor data, we propose an integrated feature selection method based on importance and relevance for descriptor data of chemical reactions.

In order to obtain more comprehensive feature information, we selected BW is used for feature screening. The features are scored according to the BW indicators, ranked according to the scores, and the top features are selected. In order to get the comprehensive and concise feature description, we screen each type of feature (molecular, atomic and vibrational descriptors) separately. Different types of descriptors are calculated and sorted by BW importance score, and those feature descriptors larger than the threshold are selected based on the appropriate threshold. And then, the selected descriptors are imported into the Random Forest [35] model for prediction. The results are shown in Figure 3A, which show that the number of feature descriptors will change with the threshold, and the prediction accuracy will also change. The experimental results show that 38 feature descriptors are selected when the threshold is 1, and the prediction result is the best (R^2 is the biggest, and RMSE is the smallest.)

From a statistical point of view, we hope to eliminate the possible correlation between features on the premise of ensuring the prediction accuracy. It is found that there are still some correlations among the 38 feature descriptors by calculating the Pearson correlation coefficient. Hence, LASSO, a classic correlation removal method in statistics is selected to further screen descriptors and get a more concise representation. By screening, 21 descriptors are reselected. Correlation visualization results for 120, 38 and 21 descriptors are shown in the heat map in Figure 3B. As can be seen from Figure 3B, the correlation between the descriptors is obviously removed after feature selection. Random forest is adopted to verify the performance of the 21 descriptors. Experiment result (R^2=0.92±0.005, RMSE=7.46±0.3) shows that the 21 descriptors can really represent the original descriptor information and still reach good prediction. Therefore, all experiments are all based on the 21 descriptors in view of simplicity and effectiveness.

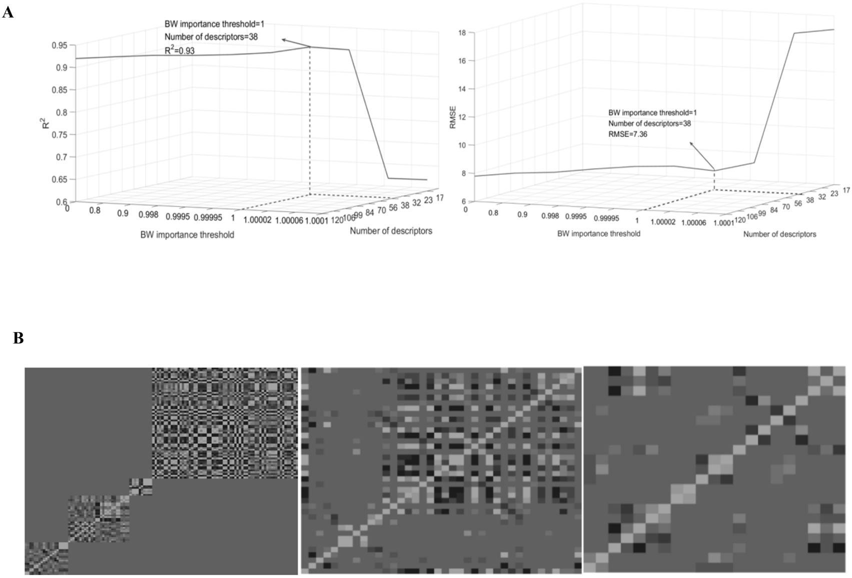
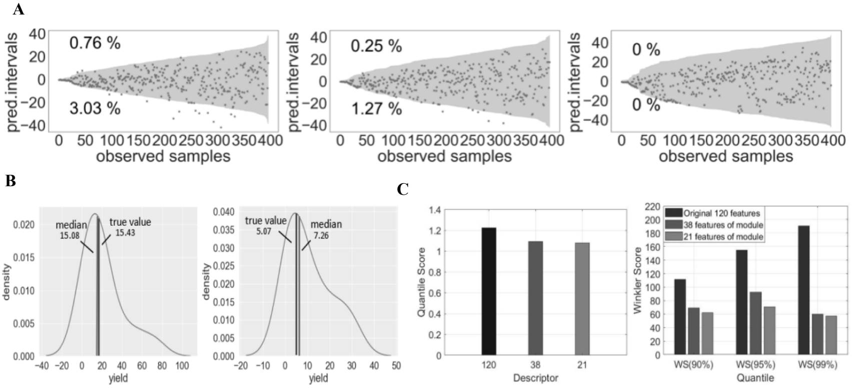


Figure 3. Feature descriptor filtering. (A) Results of random forest prediction at different thresholds of BW importance. (B) Correlation visualization results for 120, 38 and 21 descriptors (from left to right).

5.2 Yield prediction analysis based on QRF and KDE

As shown in Figure 4A, the 90%, 95% and 99% prediction intervals of the model prediction after feature screening. (In the QRF model, ten-fold cross validation is used for prediction, number of trees in the model is K=1000 and other parameters select the default value.) The prediction intervals cover all the observations with a high probability of 96.21%, 98.48% and 100%. There are two main observations: as expected, most of the forecast results are within the prediction range; the lengths of prediction intervals vary greatly. Some observations can thus be predicted much more accurately than others, indicating the accuracy of the model prediction. In addition, the interval widths of 90%, 95% and 99% prediction interval increase gradually. The main reason is that the prediction interval will gradually widen with the increase of confidence level, and the accuracy of interval estimation will gradually decrease. And vice versa, the narrower the width of the prediction interval is, the more reliable the prediction is [30]. Figure 4B shows the complete probability density curve of any two experimental yields obtained by QRF. It can be seen that the actual values are near the peak value of the curve, which indicates that the predicted values of the yield with higher probability are very similar to the actual values. The accuracy of this model is reflected.

The probabilistic prediction indicators are generally reliability indicators (prediction interval coverage) and clarity indicators (average width of forecast interval). QS and WS were used to evaluate the probabilistic prediction effect of the model in this paper. Through these two indicators, the predicted value of the original data obtained by the random forest was compared with the predicted value obtained by QRF based on the 21 descriptors obtained by re-screening, which verified the accuracy of the probabilistic prediction model proposed in this paper. Figure 4C shows the probability errors of the two models at different confidence levels. The QS values after feature screening are significantly smaller than the QS values of the original 120 features, indicating that the descriptor probability prediction effect after screening is better; at three higher confidence levels (90%, 95%, 99%), the WS values after feature screening are all smaller than those before screening, and 90% of the prediction range results are the best, so in the out-of-sample prediction experiment, we all choose 90% prediction interval.



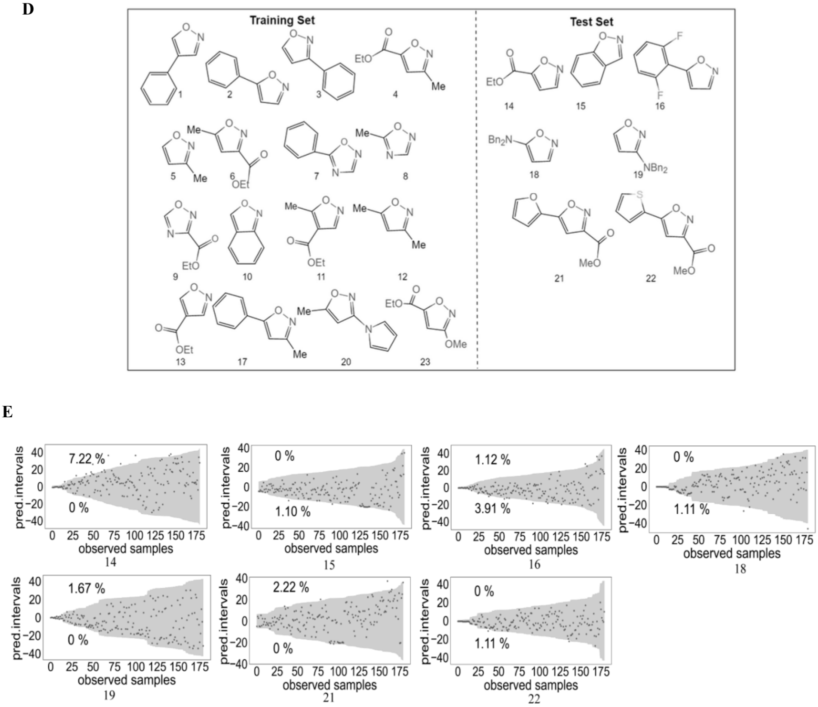


Figure 4. Predicted results. (A) 90%, 95%, 99% prediction intervals of 21 descriptors after screening. The percentage of observations above the upper bound (below the lower bound) of the different prediction intervals is shown in the upper left corner (lower left corner) of each graph. (B) Probability density estimation results (left :50 experiments, right :292 experiments). (C) Results of evaluation indicators. (Left: QS before and after feature screening, right: WS of different quantile before and after feature screening, quantile: 0.90, 0.95, 0.99). (D) Schematic illustration of the external prediction structure. (E) Out-of-sample prediction of seven additives (The randomly selected additives here are 14th, 15th, 16th, 18th, 19th, 21st, and 22nd). All kinds of additive out-of-sample prediction take 90% prediction interval as an example.

For further verify the accuracy of model prediction and implement out-of-sample prediction. Figure 4E shows the results of out-of-sample prediction. The randomly selected additives here are 14th, 15th, 16th, 18th, 19th, 21st, and 22nd, and the structure diagram is shown in Figure 4D. Above the interval, the predicted value is higher than the true value, the predicted value below the interval is lower than the true value, and only a few observations fall outside the 90% prediction interval, indicating that there is no significant systematic bias between the out of sample prediction and the model prediction, the model can predict the effect of new isoxazole or aryl halide structures on the results of Buchwald-Hartwig. It also shows that the effect of these substituents on the reaction results can be well captured by the selected 21 descriptors, and the effectiveness of the filtered descriptors is proved from the side.

QRF can be used to test outliers in addition to establishing prediction intervals to provide effective decision information. In the Figure 4A (taking the first picture 90% prediction interval as an example) there are 3 outliers above the upper boundary of the interval and 14 outliers below the lower boundary of the interval. The reaction conditions of these 17 outliers are shown in Table 1. (The numbers in the table correspond to the order of reaction types in Figure 1.)

Capturing outliers can be used to analyze data in depth. Researchers can further analyze the reaction conditions corresponding to outliers, and further experiments can be conducted to analyze whether the reflected situation is a systematic deviation or an experimental error, so as to provide more help for practical work.

Table 1. Outlier reaction conditions.

Outliers	Additive	Aryl	Base	Ligand
Above the upper boundary	19	7	1	3
	20	7	1	3
	21	12	1	4
Below the lower boundary	20	4	2	3
	20	4	3	3
	22	4	1	3
	22	4	2	3
	22	4	3	3
	19	13	3	4
	20	4	2	4
	20	13	2	1
	20	13	3	4
	21	4	1	4
	22	4	1	1
	22	7	1	4
	22	4	1	4
	22	4	2	4

When the outlier is removed, we predict the Buchwald-Hartwig amination reaction yield using Random Forest algorithm. The results as shown in Table 2, when the outlier is removed, the prediction accuracy is improved to a certain extent, both before and after the feature screening.

Table 2. Response yield prediction results before and after outlier removal.

Data	Whether to remove the outlier point	R^2	RMSE
Source data	No	0.920	7.80
	Yes	0.922	7.59
21 descriptors data	No	0.929	7.20
	Yes	0.933	7.06

By analyzing the characteristics of the 17 outlier points, we also found that: ligand_ *C7_electrostatic_charge has large values; In the reactions corresponding to the outlier point, the Aryl corresponds to the less lively halides, and the Additives are mostly concentrated in 20, 21 and 22.

6 Conclusions

QRF infer the full conditional distribution of a response variable. This information can be used to construct prediction intervals and detect outliers in the data. A quantile regression forest probability density prediction model proposed in this paper can obtain the prediction intervals of Buchwald-Hartwig amination yield under different quantile, and the probability density curve to be predicted can be obtained by probability density estimation method. It can provide effective decision information for selecting more suitable reaction conditions. Researchers can apply predictions of different quantile to the actual situation. For an unknown and harsh reaction condition, whether further experimental analysis with real compounds is needed depends on whether the prediction results are within the prediction range under the prediction of higher quantile, if in, further chemical analysis can be carried out. Thus, it is possible to efficiently determine whether an unknown reaction condition can be used for further analysis. The detection of outliers and the effective out of sample prediction provide more possibilities for researchers in practical applications.

Acknowledgement: The authors would like to gratefully thank the Key Scientific Research Projects of Colleges and Universities in Henan Province (grant no. 21A150014) for funding and support.

References

[1] E. J. Corey, W. T. Wipke, R. D. Cramer, W. J. Howe, Techniques for perception by a computer of synthetically significant structural features in complex molecules, *J. Am. Chem. Soc.* **94** (1972) 431–439.

[2] J. B. Hendrickson, Descriptions of reactions: Their logic and applications, *Recl. Trav. Chim. Pays-Bas.* **111** (1992) 323–334.

[3] 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.

[4] 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.

[5] M. Fujinami, J. Seino, T. Nukazawa, S. Ishida, T. Iwamoto, H. Nakai, 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.

[6] 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.

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

[8] M. Fujinami, H. Maekawara, R. Isshiki, J. Seino, J, Yamaguchi, H. Nakai, Solvent selection scheme using machine learning based on physicochemical description of solvent molecules: application to cyclic organometallic reaction, *Bull. Chem. Soc. Jpn.* **93** (2020) 841–845.

[9] 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.

[10] S. Ekins, The next era: deep learning in pharmaceutical research, *Pharm Res.* **33** (2016) 2594–2603.

[11] 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.* **5** (2019) 4275–4282.

[12] M. H. S. Segler, M. Preuss, M. P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, *Nature* **555** (2018) 604–610.

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

[14] A. C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen, T. F. Jamison, Reconfigurable system for automated optimization of diverse chemical reactions, *Science* **361** (2018) 1220–1225.

[15] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, Accelerated search for materials with targeted properties by adaptive design, *Nat. Commun.* **7** (2016) 11241–11249.

[16] R. F. Heck, J. P. Nolley, Heck reaction, *J. Org. Chem.* **37** (1972) 2320–2322.

[17] E. Negishi, A. O. King, N. Okukado, Selective carbon-carbon bond formation via transition metal catalysis. 3. A highly selective synthesis of unsymmetrical biaryls and diarylmethanes by the nickel- or palladium-catalyzed reaction of aryl- and benzylzinc derivatives with aryl halides , *J. Org. Chem.* **42** (1977) 1821–1823.

[18] A. O. King, N. Okukado, E, Negishi, Highly general stereo-, regio-, and chemo-selective synthesis of terminal and internal conjugated enynes by the Pd-catalysed reaction of alkynylzinc reagents with alkenyl halides, *J. Chem. Soc. Chem. Commun.* **19** (1977) 683–684.

[19] N. Miyaura, K. Yamada, A new stereospecific cross-coupling by the palladium-catalyzed reaction of 1-alkenylboranes with 1-alkenyl or 1-alkynyl halides, A. Suzuki, *Tetrahedron Lett.* **20** (1979) 3437–3440.

[20] N. Miyaura, A. Suzuki, Stereoselective synthesis of arylated (E)-alkenes by the reaction of alk-1-enylboranes with aryl halides in the presence of palladium catalyst, *J. Chem. Soc. Chem. Commun*. **19** (1979) 866–867.

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

[22] J. F. Hartwig, Evolution of a fourth generation catalyst for the amination and thioetherification of aryl halides, *Acc. Chem. Res.* **41** (2008) 1534–1544.

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

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

[25] P. S. Kutchukian, J. F.Dropinski, K. D. Dykstra, B. Li, D. A. DiRocco, E. C. Streckfuss, L. C. Campeau, T. Cernak, P. Vachal, I. W. Davies, S. W. Krska, S. D. Dreher, Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods, *Chem. Sci.* **7** (2016) 2604–2613.

[26] E. Vitaku, D. T. Smith, J. T. Njardarson, Analysis of the structural diversity, substitution patterns, and frequency of nitrogen heterocycles among U.S. FDA approved pharmaceuticals, *J. Med. Chem.* **57** (2014) 10257–10274.

[27] J. P. Jia, Q. X. He, Y. Jin, *Statistics*, China Renmin Univ. Press, Renmin, 2009.

[28] S. Dudoit, J. Fridlyand, T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *J. Am. Stat. Assoc.* **97** (2002) 77–87.

[29] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Statist. Soc. B.* **58** (1996) 267–288.

[30] N. Meinshausen, G. Ridgeway, Quantile regression forests, *J. Mach. Learn. Res.***7** (2006) 983–999.

[31] R. Koenker, G. Bassett Jr, Regression quantiles, *Econometrica* **46** (1978) 33–50.

[32] Y. Song, L. Wang, Y. Liu, Y. Zhang, Q. Yang, C. Ma, H. Ma, Fitting method for probability distribution function based on nuclear density estimation, *PSCE* **32** (2016) 85–88.

[33] Y. Wang, D. Gan, M. Sun, N. Zhang, Z. Lu, Probabilistic individual load forecasting using pinball loss guided LSTM, *Appl. Energy* **235** (2019) 10–20.

[34] G. Li, D. C. Wu, M. Zhou, A. Liu, The combination of interval forecasts in tourism, *Ann. Tour. Res.* **75** (2019) 363–378

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