**MATCH**
*Communications in Mathematical and in Computer Chemistry*

*MATCH Commun. Math. Comput. Chem.* **91** (2024) 41–78
**ISSN:** 0340–6253
**doi:** 10.46793/match.91-1.041W

# ChemCNet: An Explainable Integrated Model for Intelligent Analyzing Chemistry Synthesis Reactions

**Lanfeng Wang[a], Hengzhe Wang[b], Shuoshi Liu[b], Zixin Li[b], Yaping Yu[b], Yun Chai[c,*], Xiaohui Yang[b,*]**

[a] *School of Mathematics and Statistics, Anyang Normal Unieversity, Anyang, China, 455000*

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

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

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

(Received March 15, 2023)

**Abstract**

Palladium (Pd)-catalyzed cross coupling reactions are of great significance in organic synthesis. However, the reaction route is more complex, time-consuming and costly. For addressing the above problems, a model-related feature selection strategy is introduced, focusing on iterative optimization of feature description and prediction to guide and strengthen each other. Then, we combine the lightweight convolution neural network (CNN) driven by attention mechanism with CatBoost to build an intelligent chemical synthesis reaction analysis model—ChemCNet. Moreover, we conduct the interpretability analysis based on ChemCNet model. The results show that ChemCNet model has achieved relatively high prediction accuracy and generalization, and it is helpful to provide reliable decision-making information for the experimenter or institution.

*Corresponding author.

# 1 Introduction

Since the amount of valuable natural products in nature is very small and cannot satisfy the demand, the artificial synthesis of natural products has become very important, and synthetic methodology has become the most important part of the field of organic chemistry at home and abroad. Among the organic synthesis reactions, the Buchwald-Hartwig coupling reaction is the most advanced C-N coupling reaction available. Richard F. Heck discovered in the 1970s that the linkage between carbon atoms could be achieved under milder conditions using a palladium catalyst [1], Subsequently, Ei-ichi Negishi and Akira Suzuki further developed the method of using palladium-catalyzed carbon-carbon bond cross-coupling [2–5]. In 2010, R. F. Heck, Ei-ichi Negishi and A. Suzuki were awarded the Nobel Prize in Chemistry for the development of "Palladium-catalyzed cross-coupling methods in organic synthesis". In 2022, C. R. Bertozzi, M. Meldal, and K. B. Sharpless were awarded the Nobel Prize in Chemistry for their significant contributions to "developments in click chemistry and bioorthogonal chemistry". The creation of these coupling synthesis methods has enabled chemists to manipulate atoms and molecules in an unprecedented degree.

Currently, automated organic synthesis frees people from complex, dangerous, and boring working environments, increasing efficiency and precision. With theoretical modeling and technological innovations, complex chemical reactions can be simulated, synthesized with the high-speed processing power of computers, and the cross-fertilization of artificial intelligence algorithms with chemical disciplines is of great importance to advance academic research [6]. In 2018, D. T. Ahneman et al. [7] reported the prediction of the yield of the Buchwald-Hartwig amination reaction by random forests, an advanced study of machine learning methods in the field of multidimensional chemical space prediction; M. H. S. Segler et al. [8] proposed the use of recurrent neural networks as a generative model for molecular structures; J. Dong et al. [9] used the XGBoost model as a prediction model, and X. H. Mu et al. [10] used Deep Forest as a model, both of which improved the prediction accuracy. However, these works

are not strong enough for deep feature mining of data, and further feature learning is a direction worth thinking about.

Obtaining good features is the key to successful recognition in machine learing , and finding good data representations is the core task of machine learning. In general, key information exists in only a small number of features, so only a small number of key features are sufficient to provide enough information. Therefore, it is necessary to obtain comprehensive and clean data to improve the prediction performance of the model. J. Dong et al. [9] proposed a feature selection method based on importance and relevance to successfully reduce the feature dimension, and X. H. Yang et al. [11] constructed a gene selection method based on decision information factor (DIF). M. Chieregato et al. [12] adopt an all relevant feature selection method to select features. Features are closely related to the model, we hope that the selected features can improve the effect of the model. Therefore, how to select relevant features based on the model is interesting and necessary.

Deep learning is the most flexible representation learning that extracts highly abstracted feature through layer-by-layer networks, which is better able to portray the rich intrinsic information of data. Deep learning shines in a variety of tasks and has shown great potential in the last few decades. 2019, C. Coley et al. [13] proposed the use of graph convolutional neural network models to predict organic reaction products given reactants, reagents, and solvents. H. X. Hou et al. [14] proposed a one-dimensional CNN with added attention mechanism to predict reaction yields; 2021, Y. N. Zhao et al. [15] used Deep Convolutional Neural Networks to predict reaction yields. Z. T. Song et al. [16] proposed an attention-based multi-label neural networks for integrated prediction.

Motivated by these works, a feature selection method related to the model is first used to screen features. On this basis, an intelligent and lightweight prediction system is proposed, which combines a lightweight attention driven CNN and the integration tree CatBoost [19] to integrate feature representation learning and regression prediction into a model, guiding and reinforcing each other, and obtains more desirable prediction results.

The main contributions of this paper are as follows.

(1) The model-related feature selection method for chemical descriptors can not only reduce data dimension, but also enhance the expression effect of the model.

(2) Theoretically, ChemCNet model integrates a lightweight convolutional neural network and tree model into one model, and adds a lightweight attention mechanism to focus on key features.

(3) Structure analysis, feature learning performance, prediction performance and generalization performance and interpretability analysis (feature importance,ALE value [17] and SHAP value [18]) of ChemCNet model are studied.

# 2 Intelligent prediction model for reaction yield – ChemCNet

## 2.1 Model-based feature selection

In order to select a subset suitable of the features, RFE [20,21] (Recursive Feature Elimination, RFE) backward search method and SHAP value as feature evaluation criteria are used for feature selection.

Firstly, according to the principle of backward search, the features with the lowest feature score will be removed, and then the model will continue to be constructed on the remaining features to regain a new round of feature ranking, and then the features with the lowest feature score will be removed, and this process will be repeated to sequentially remove the features with the lowest score until the specified number of features is reached. In each iteration, the current set of remaining features is re-evaluated, and the score of each feature is adjusted in the iterative process, and finally presented in the form of RMSE (Root Mean Square Error, RFE). The specific principle is as follows.

Start with an empty set of eliminated features $E = \{\}$. Calculate the

current loss value:

$$L_{-E} = \sum_{i=1}^{N} L_{i,\{-E\}} = \sum_{i=1}^{N} l(y_i, a_i - \sum_{k \in E} v_{i,k}). \tag{1}$$

For each available feature, use the SHAP value to calculate the score when the loss function changes:

$$\begin{aligned} score_j &= L_{\{-E,-j\}} - L_{-E} \\ &= \sum_{i=1}^{N} l(y_i, a_i - \sum_{k \in E} v_{i,k} - v_{i,j}) - L_{-E}. \end{aligned} \tag{2}$$

Remove a feature with the lowest score and add it to the set. If the feature still needs to be eliminated, repeat this process.

## 2.2 ChemCNet

Based on the data, this paper designs a lightweight convolutional neural network and adds a lightweight attention module to focus on the key features without significantly increasing the complexity of the model. And, in order to avoid the over-fitting risk brought by the full connection layer and improve the prediction efficiency, the last full connection layer is replaced by CatBoost, finally a hybrid model ChemCNet is built.

### 2.2.1 Feature representation learning

The classical DCNN model includes convolution layer, pooling layer, activation function and full connection layer. Among them,each convolutional kernel can be considered as a feature filter to extract important features, and the pooling layer has the characteristic of compression, and a large amount of redundant information has been removed in subsection 2.1. Taking into account the chemical background of the data, it is considered to remove the pooling layer. Therefore, by simply stacking the input layer, hidden layer and output layer, a network feature representation learning model is established. The implicit layer is composed of four convolution layers (the activation function of each layer is Rectified Linear Unit) and

three full connection layers.

### 2.2.2 Attention mechanism

To focus on key features and enhance the expressiveness of the model, an attention mechanism layer is further added. Set ECA [22] module as the attention mechanism layer, a local cross-channel interaction strategy that does not require dimensionality reduction, which effectively captures the information of cross-channel interactions and can be efficiently implemented by one-dimensional convolution. It is a lightweight attention module with simple operation.

Let the output of a convolutional block be $X \in R^{W \times H \times C}$, where W, H, and C are the width, height, and channel dimension (number of channels) (i.e., the number of filters), aggregated features $y \in R^C$ without dimension reduction. The ECA module uses a band matrix to learn channel attention:

$$\begin{bmatrix} w^{1,1} & \cdots & w^{1,k} & 0 & 0 & \cdots & \cdots & 0 \\ 0 & w^{2,2} & \cdots & w^{2,k+1} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & w^{C,C-k+1} & \cdots & w^{C,C} \end{bmatrix}. \tag{3}$$

The $k \times C$ parameter is involved, and the complete independence between different groups in the equation is avoided. The calculation of the weight $y_i$ only considers the interaction with its $k$ neighbors, i. e. $\omega_i = \sigma(\sum_{j=1}^{k} w_i^j y_i^j), y_i^j \in \Omega_i^k$, where $\Omega_i^k$ represents the set of $k$ adjacent channels. A more efficient approach is to make all channels share the same learning parameters, i. e.,

$$\omega_i = \sigma\left(\sum_{j=1}^{k} w^j y_i^j\right), y_i^j \in \Omega_i^k, \tag{4}$$

Note that this strategy can be easily implemented by a fast 1D convolution with kernel size $k$, i. e.,

$$\omega = \sigma(C1D_k(y)), \tag{5}$$

where C1D represents one-dimensional convolution. This is the ECA module, which only involves $k$ parameters. The range $k$ of interaction coverage (i.e., the kernel size $k$ of one-dimensional convolution) is determined according to formulas (6) and (7).

$$C = \phi(k) = 2^{(\gamma * k - b)}, \tag{6}$$

given the channel dimension $C$, the kernel size $k$ can be adaptively determined:

$$k = \varphi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd}, \tag{7}$$

where $\gamma$ is the coefficient of the primary term, $b$ is the constant term, $| t |$ odd represents the nearest odd number of $t$. Obviously, by mapping $\varphi$, high-dimensional channels have longer interactions, while low-dimensional channels have shorter interactions by using nonlinear mapping. Set according to data characteristics $b = 1$,$\gamma = 2$.

### 2.2.3 ChemCNet prediction model

Because the traditional connection layer is easy to be excessive, and the good interpretability helps the experimenter to better optimize the design route and provide constructive suggestions, we replace the final output layer with CatBoost, and import the features extracted from the last full connection layer into the CatBoost model as new input data for training and prediction. ChemCNet prediction model has excellent feature learning ability, ability to focus on key features and good interpretability of tree model. CatBoost built the final model $F^T = \sum_{t=1}^{T} f^t$ by integrating the weak learner $f^t$. The model loss function is set as:

$$L(f(x), y) = \sum_i w_i \cdot l(f(x_i), y_i) + J(f), \tag{8}$$

where $L(f(x), y)$ is the loss value at point $(x, y)$, $w_i$ is the weight of $x_i$, and $J(f)$ is the regular term.

In the training tree building process, the trees are constructed sequentially and the goal of the next tree is to fit the negative gradient

$g_i = -\frac{\partial l(a,g_i)}{\partial a}\Big|_{a=F^{T-1}(x_i)}$ of the loss function $l$, where $a_i = f(x_i)$, $w_i$ is the weight of $x_i$. Therefore, the gradient descent method is used to optimize the loss function. Grading function $Score(a,g) = S(a,g)$ is needed to measure the quality of gradient fitting. CatBoost implements both a first-order gradient version and a version of XGBoost [23] Taylor expansion that introduces a second-order gradient, as well as extending some other scoring functions to determine whether the leaves are split (as shown in Eq.9), and CatBoost allows the user to freely choose whether to use a first-order gradient or a second-order gradient. ChemCNet model uses the L2 scoring function.

$$\begin{aligned} L2 &= -\sum_i w_i (a_i - g_i)^2, \\ Co\sin e &= \frac{\sum\limits_i w_i \cdot a_i \cdot g_i}{\sqrt{\sum\limits_i w_i {a_i}^2} \cdot \sqrt{\sum\limits_i w_i {a_i}^2}}. \end{aligned} \tag{9}$$

Finding the optimal tree structure is an iterative process, and for the sake of explanation, assume that the depth of the tree to be constructed is 1. The structure of a tree like this needs to be determined by the index $j$ of some features and the boundary value $c$. Let $x_{ij}$ denote the $j$-th feature of the $i$-th sample and $a_{left}$,$a_{right}$, denote the left and right leaf nodes of the tree, respectively. When $x_{ij} \le c, f(x_i) = a_{left}$, when $x_{ij} > c, f(x_i) = a_{right}$. So now the goal is to find the optimal $j$ and $c$ with the help of scoring function, so that the optimal tree structure is found. Thus we have:

$$\begin{aligned} S(a,g) &= -\sum_i w_i (a_i - g_i)^2 \\ &= -\left( \sum_{i:x_{ij} \le c} w_i (a_{left} - g_i)^2 + \sum_{i:x_{ij} > c} w_i (a_{right} - g_i)^2 \right). \end{aligned} \tag{10}$$

Let $W_{left} = \sum\limits_{i:x_{ij} \le c} w_i, W_{right} = \sum\limits_{i:x_{ij} > c} w_i$, by deriving the optimal values of $a_{left}$,$a_{right}$, are: $a^*_{left} = \frac{\sum\limits_{i:x_{ij} \le c} w_i g_i}{W_{left}}$, $a^*_{right} = \frac{\sum\limits_{i:x_{ij} > c} w_i g_i}{W_{right}}$, which

are brought back to the scoring function, and the following equation can be obtained after expanding the brackets and removing the constant term.

$$j^*, c^* = \arg\max_{j,c} W_{left} \cdot (a^*_{left})^2 + W_{right} \cdot (a^*_{right})^2. \tag{11}$$

When the depth is greater than 1, the scoring function will change to $Score(a, g) = \sum_{leaf} S(a_{leaf}, g_{leaf})$, at which point the following equation is available.

$$j^*, c^* = \arg\max_{j,c} S(\overline{a}, g), \tag{12}$$

where $\overline{a}$ is the best leaf value obtained after the partitioning of $j$ and $c$.

On the other hand, CatBoost introduces an "artificial timeline", a timeline based on the arrival of training examples so that only "previously seen" examples can be used when calculating statistics. That is, for each sample $X_k$, a separate model $M_k$ is trained from the training set that does not contain sample $X_k$. The model is used to estimate the gradient and use this estimate to score the resulting tree. That is, only the current model trained on the previous samples is used to update the gradient of the new samples of the model, which provides unbiased gradients and effectively avoids the overfitting problem caused by biased pointwise gradient estimation that is common to all classical boosting algorithms. In each step $t$ of the learning process, each model can be interpreted as an approximation of a model $F^t$. For each permutation $\sigma$, n different models $M_i$ need to be trained, and for each model $M_i$, $M_i(X_1), ..., M_i(X_n)$ must be updated. Therefore, the resulting complexity of this operation is $O(sn^2)$. This increases memory consumption and time complexity, so CatBoost is chosen to maintain $log_2 n$ models: $M_i^{'}(X_j), i = 1, ..., [log_2 n], j < 2^{i+1}$, where $M_i^{'}(X_j)$ is an approximation of the previous $2^i$ samples based on sample $j$. Then, the number of predictions $M_i^{'}(X_j)$ will be no greater than $\sum_{0 \le i \le \log_2 n} 2^{i+1} < 4n$. This operation reduces the complexity of a tree construction to $O(sn)$. Thus, CatBoost first uses unbiased estimation of gradient step size to select tree structure, and then performs standard GBDT(Gradient Boosting Decision Tree). First initialize an empty tree $T$, find all possible splitting methods using the greedy algorithm, form a

candidate splitting set $C$, select any one split $c \in C$ from $C$, assign the split $c$ to the tree $T$, noted as $T_c$, calculate the value of the leaf node as $leaf_i = GetLeaf(x_i, T_c, \sigma), i = 1, 2, ..., n$, calculate the average of the gradient of the $i-th$ leaf nodes as $\Delta_i = avg(grad_{\lfloor \log_2(\sigma(i)-1) \rfloor}(p)x_j, r_j, )$, where $p : leaf_p = j, \sigma(p) < \sigma(i), i = 1, 2, ..., n$, calculate the value of the loss function $Loss(T_c) = \|\Delta - grad\|_2$ corresponding to the $c-th$ split, select the smallest loss function corresponding to $T = \arg\min_{T_c}(loss(T_c))$, which is the final output.

In summary, the workflow of ChemCNet prediction model is as follows.

(1) The dataset is divided into training and testing sets according to 7:3

(2) The obtained 24-dimensional feature data are jointly obtained as 25-dimensional data and normalized.

(3) Model training. The network loss function is set as MSE (Mean Square Error), and the optimization algorithm is Adam. When the loss function converges to the smallest value, the parameters of the network model are saved. The feature data extracted from the third fully connected layer is used as training data and imported into the CatBoost model for training.

(4) Model testing. Similar to (3), feature extraction is performed on the test set, and then the extracted features are imported into CatBoost to predict the reaction yield and fit analysis with the actual reaction yield to give the accuracy of the prediction.

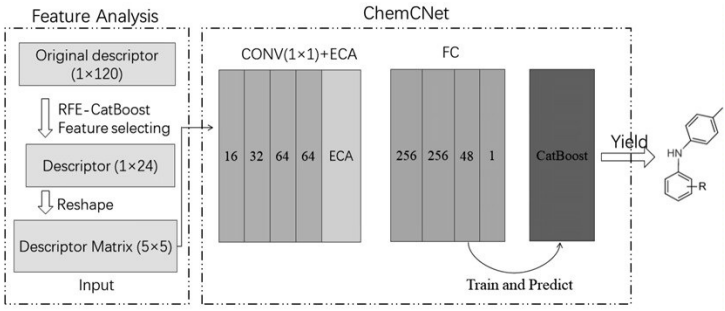



**Figure 1.** Flow chart of ChemCNet prediction model.

# 3 Experimental results

This section analyzes ChemCNet from feature analysis, structure analysis, feature learning performance, prediction performance and generalization performance. Then, feature importance, ALE value and SHAP value are used to explore the internal relationship between reaction conditions and reaction yield, so as to provide more decision-making information for the experimenter.

The machine learning work involved is implemented by a combination of the Scikit-learn package (version 0.24.2) in Python (version 3.6.13), Tensorflow 1.15.

## 3.1 Data description

We select the data on Buchwald-Hartwig coupling reaction published by Ahneman et al. Ahneman et al. used an ultra-high-throughput device for coupling reactions and obtained data for 4608 reactions (including controls) spanning different reaction combinations consisting of 4 components, including 23 isoxazole additives, 15 aryl or Heteroaryl halide, 4 palladium catalyst ligands and 3 bases. The yields of these reactions are used as the model output. The effective experimental data are 3960. According to the characteristics of chemical reactions, 120 descriptors of related chemical properties are selected, including electronegativity, dipole moment, NMR shift, energy of frontier molecular orbital and so on. Chemical descriptors of reactants, catalysts, and additives involved in the reaction are independent variables, and the corresponding reaction yields are dependent variables. Correspondingly, in order to avoid time-consuming analysis and logging of computational data, Doyle et al. developed software to submit molecular, atomic, and vibrational property calculations to Spartan and subsequently extract these features from the resulting text files for accessibility to a general user. The program requires only the input of reagent structures in the Spartan graphical user interface and specification of the reaction components in a Python script; it is applicable to any reaction type. The program then generates the data table that can be used for modeling. Therefore, the 120 descriptors can be divided into three cate-

gories: molecular descriptors (28), atomic descriptors (64) and vibrational descriptors (28) .

## 3.2 Feature analysis

The advantage of RFE is that the number of features can be set independently, which gives more flexibility and choice. The 60, 50, 45, 40, 35, 30, 25, 20, 15, 10, and 5 feature descriptors were selected for experiments.

The feature selection process is shown in Fig.2. X-axis is the number of features and y-axis is the loss function value RMSE of the model. Left 1 shows that when the feature descriptors are 60 (RMSE is 5.559), 40 (RMSE is 5.566), 25 (RMSE is 5.684), and 15 (RMSE is 5.651), the RMSE of the model is the smallest. Further narrowing the search interval (left 2, left 3) reveals that the model has the smallest RMSE of 5.645 when 24 feature descriptors are used for prediction. Although the RMSE is still higher than the results of 60 and 40 feature descriptors, 24 feature descriptors are finally selected as the features, considering that the number of 40 and 60 feature descriptors is still high and the operation is more complicated and time-consuming. The correlation heat map (Fig.2.(b) ) shows that the relevance among the features after feature filtering is obviously removed relevance with the original data of 120 feature descriptors, and a concise and comprehensive feature descriptor data is obtained, which will provide convenient input and training for the subsequent model.

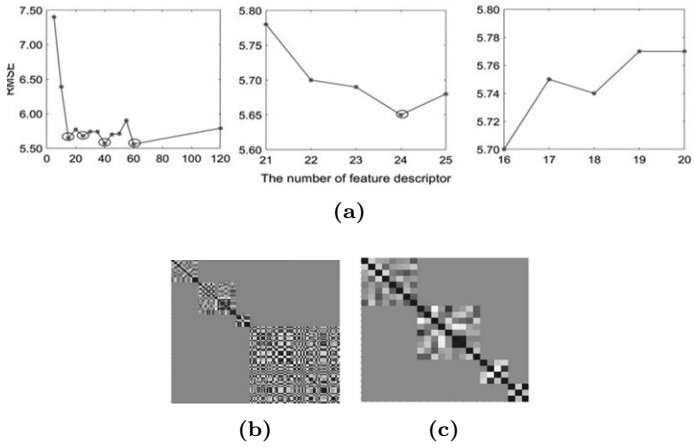



**Figure 2.** Feature analysis. (a)Feature descriptor screening process. (b)Relevance heat map visualization results for the original 120 descriptors. (c)Relevance heat map visualization results of the 24 feature descriptors obtained after feature screening.

## 3.3 Structure analysis

The structure of the network without CatBoost is denoted as ChemCNet-0. Generally speaking, the more convolutional layers, the better the feature learning ability of the network and the more comprehensive the information extracted. However, the increase of convolution layers brings more parameters, which increases the complexity of the model and the difficulty of training. Therefore, the fully connected layer is first fixed to 2 layers first, and then the effect of 2-4 convolutional layers on the prediction results is analyzed. The experimental results are shown in Fig.3.(a) , it can be seen that RMSE decreases and $R^2$ increases with the increase of the number of convolutional layers. When the number of convolution layers is 5, the model appears over-fitting, and the prediction results begin to decline. When the number of convolution layers is 4, the model reaches the optimal value. Therefore, setting the number of convolution layers to 4 can better extract the information of feature descriptors and obtain better prediction results. The analysis results of the full connection layers are

the same (experimental results are shown in Fig.3.(b)), when the number of full connection layers is 4, the best result is obtained.

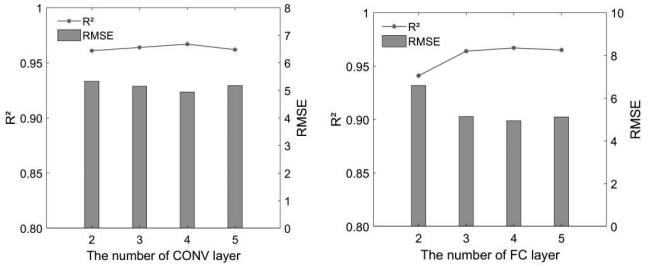



**Figure 3.** Structure analysis. (a) The influence of different convolution layers on the prediction results. (b) The influence of different full connection layers on the prediction results.

Since the feature dimension is low, the kernel size is set to (2,2), so at most two pooling layers can be added. The number and location of pool layers are analyzed as follows. The experimental results are shown in Table 1. It can be seen from Table 1 that when one layer of pooling layer is added, the prediction accuracy of the network is not ideal. After two layers of pooling layer are added, the prediction accuracy is worse, indicating that adding pooling layer is not suitable for the data in this paper, which will seriously affect the learning ability of the network. The more layers are added, the worse the prediction accuracy of the network. Therefore, this paper chooses to remove the pool layer.

Then, the location of ECA module is analyzed to select the optimal location. The parameter r and b are set to 2 and 1 respectively. The ECA module is placed after the first convolution, the second convolution, the third convolution and the fourth convolution respectively. The experimental results are shown in Table 2. It can be seen that after the ECA module is placed in the fourth layer of convolution, the prediction result of the model is the best, which can improve the network performance without significantly increasing the network complexity.

In summary, ChemCNet-0 contains 4 convolutional layers, 4 fully connected layers and one attention mechanism layer. The nodes of the con-

volutional layer are 16, 32, 64, 64, respectively , and the kernel size and step size are 1×1, and the nodes in the fully connected layer is 256, 256, 48, and 1 (output layer), respectively.

**Table 1.** Impact of Pooling Layer on prediction Results.

| Location of pooling layer | $R^2$ | RMSE |
|---|---|---|
| After the first CONV layer | 0.57 | 17.76 |
| After the second CONV layer | 0.64 | 16.27 |
| After the third CONV layer | 0.65 | 16.06 |
| After the forth CONV layer | 0.66 | 15.9 |
| **Average** | **0.63** | **16.5** |
| Place one layer after the first and second CONV layers | 0.37 | 21.41 |
| Place one layer after the first and third CONV layers | 0.51 | 19.17 |
| Place one layer after the first and forth CONV layers | 0.56 | 17.95 |
| Place one layer after the second and third CONV layers | 0.61 | 16.99 |
| Place one layer after the second and forth CONV layers | 0.62 | 16.79 |
| Place one layer after the third and forth CONV layers | 0.64 | 16.3 |
| **Average** | **0.55** | **18.1** |

**Table 2.** Prediction results of ECA module at different positions of ChemCNet-0.

| ECA module location | $R^2$ | RMSE |
|---|---|---|
| After the first CONV layer | 0.9646 | 5.07 |
| After the second CONV layer | 0.9651 | 5.06 |
| After the third CONV layer | 0.9645 | 5.12 |
| After the forth CONV layer | **0.9669** | **4.91** |

## 3.4 Feature learning performance analysis

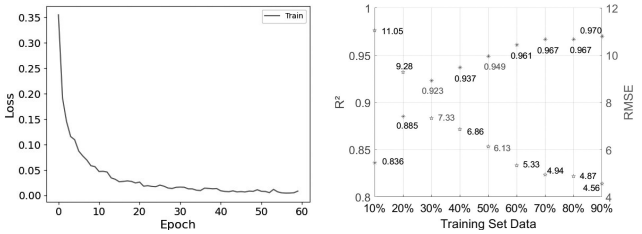
The framework of ChemCNet-0 model has been built, the convergence analysis is then performed. The convergence curves are shown in Fig.4. From Fig.4.(a), it can be seen that the loss value (Loss) decreases and eventually plateaus as the number of iterations increases, indicating that ChemCNet-0 is convergent and can be used for the next experiments.

We find that relatively small training sets can also effectively explore hidden feature information, and play a good learning performance. As shown in Fig.4.(b), 10% - 90% of the total data sets are selected as training sets. It can be seen from the figure that the value of $R^2$ is 0.923, and the value of RMSE is 7.33 when the training set only accounts for 30% of the data set, which is better than the results in Ref. [7]. The value of $R^2$ is 0.949, and the RMSE is 6.13. When the training set accounts for 50% of the data set. This result indicates that ChemCNet-0 has acceptable accuracy to predict the yield by using relatively smaller training set (50%) compare with the conventional 70/30 split of dataset.

Comparing with ChemNet [15] and LetNet [15], ChemCNet-0 has better learning ability and the prediction accuracy reaches $R^2 = 0.97$ and RMSE=4.93 (Fig.4.(c)), which shows the applicability and superiority of ChemCNet-0 to accurately and comprehensively capture the important feature information, and the simpler structure of ChemCNet-0 effectively reduce the model complexity.

In addition, in order to test the generalization ability of ChemCNet-0, we conduct 10 times of 10 fold cross validation. The schematic diagram of cross validation is shown in Fig.4.(d), and the experimental results of cross validation are shown in Fig.4.(e). The experimental results show that the feature learning module ChemCNet-0 can not only obtain more abstract data features, but also accurately predict the reaction yield, and has a good generalization ability.

In summary, ChemCNet-0 has excellent feature learning ability to further acquire deeper features hidden inside the data, and also has good generalization performance.




| Trial | $R^2$ | RMSE |
|---|---|---|
| 1 | 0.969 | 4.73 |
| 2 | 0.970 | 4.69 |
| 3 | 0.968 | 4.92 |
| 4 | 0.969 | 4.89 |
| 5 | 0.970 | 4.65 |
| 6 | 0.959 | 5.49 |
| 7 | 0.973 | 4.54 |
| 8 | 0.972 | 4.64 |
| 9 | 0.974 | 4.50 |
| 10 | 0.959 | 5.50 |
| Average Result | 0.968 | 4.86 |

(e)

**Figure 4.** Feature learning performance analysis of ChemCNet model. (a) Convergence curve of ChemCNet-0. (b) prediction performance of ChemCNet-0 under different training sets. (c) Comparison of prediction accuracy of LetNet, ChemNet and ChemCNet-0. (d) Cross validation method and data set partition. (e) The experimental results of ChemCNet-0 10 fold cross validation method.

## 3.5 Prediction performance analysis

The depth features extracted from ChemCNet-0 are imported into CatBoost as new input data for training and prediction. In order to better fit

the real data, we used ten fold cross validation and grid search method to obtain the optimal parameters of CatBoost model.

Then, Student's t test is performed on the actual and predicted values, the results are shown in Fig.5.(a). There is no significant difference between them, and the prediction results are accurate. We compare the prediction accuracy of ChemCNet with CatBoost and other machine learning methods, including GBDT, Random Forest (RF), Decision Tree (DT), AdaBoost (Ada), k-nearest neighbor (KNN), Ridge regression (Rid), Linear (Lin), and Extreme Random Tree (Extra), and $R^2$ and RMSE are used as evaluation metrics. The prediction results are shown in Fig.5.(b). It is easy to see that the prediction accuracy of the Linear Regression is low and not suitable for application to this chemical reaction data; Although Decision Tree and other machine learning methods have improved the prediction accuracy, the results are still unsatisfactory. CatBoost has substantially improved over these methods, but ChemCNet prediction model has better prediction performance and better fit to the real yield with $R^2 = 0.97$ and RMSE=4.88.

In order to make the results clearer, we use residual to analyze the error of CatBoost and ChemCNet regression prediction. In the residuals plot, if the residuals are evenly distributed within a horizontal strip with a residual of about 0, the selected model has a higher degree of fit, and the narrower the strip is, the higher the fitting accuracy. Fig.5.(c) shows the residual plots of the CatBoost and ChemCNet prediction models, respectively. CatBoost has a larger residual and wider distribution area, while the residuals of the ChemCNet are more concentrated around the straight line y=0, with a more concentrated distribution and narrower distribution area, and the prediction effect is significantly better than that of CatBoost, which again proves the excellent prediction performance of the ChemCNet model.

Fig.5.(c) shows the box plots of the prediction accuracy (RMSE and $R^2$) of the above models, from which it can be seen that the ChemCNet has the smallest boxes, indicating that the ChemCNet has the most concentrated data distribution of the experimental results, the least data volatility, and the best and most stable prediction performance. From Fig.6.(a),

it can be seen that the ChemCNet can also obtain acceptable prediction accuracy compare to CatBoost in the case of small samples.Moreover, as shown in Fig.6.(b), the ChemCNet also achieves higher prediction accuracy under sparse data, and the prediction results are better than those of the CatBoost.

While keeping the ChemCNet-0 the same, the ChemCNet model is compare with the ChemCNet-0+GBDT, ChemCNet-0+Decision Tree, Che-mCNet-0+Random Forest, and ChemCNet-0+Extra Tree hybrid models, respectively, and from Fig.6.(c), it can be found that the prediction accuracy of the ChemCNet-0+machine learning model is better than single machine learning regressor, and among them, ChemCNet has the lowest RMSE value, which indicates that the model has better prediction performance. It can also be seen that the prediction accuracy of the models (GBDT, Decision Tree, Random Forest, and Extra Tree) are significantly improved after the network feature learning, which again proves the excellent feature extraction ability of ChemCNet-0.

In addition, ablation experiments were also conducted (Table 4-4): comparison of lines 1, 2 and 4 shows that the prediction accuracy of the single model (CatBoost, ChemNet-0) is not as good as that of the hybrid model ChemCNet; The comparison of rows 1 and 3 shows that adding feature representation learning can effectively improve the prediction accuracy; The comparison between lines 2 and 4 shows that the prediction accuracy can be improved again after replacing the full connection layer with CatBoost; The comparison of 3 and 4 lines shows that the addition of attention mechanism layer has not significantly increased the complexity of the model, but also played a certain role in improving the prediction results.

It can be seen from Fig.5 that ChemCNet can also obtain acceptable prediction accuracy compare with CatBoost in the case of small samples. In addition, the ablation study (Table 1) show that the prediction accuracy of the single model (CatBoost, ChemNet-0) is inferior to that of the hybrid model ChemCNet; the comparisons in rows 1,3 show that adding feature representation learning can effectively improve the prediction accuracy. The comparisons in rows 1,4 show that adding attention-driven feature

learning improves the prediction accuracy again; the comparisons in rows 3,4 show that adding an attention mechanism layer helps to improve the prediction results and does not significantly increase the complexity of the model.

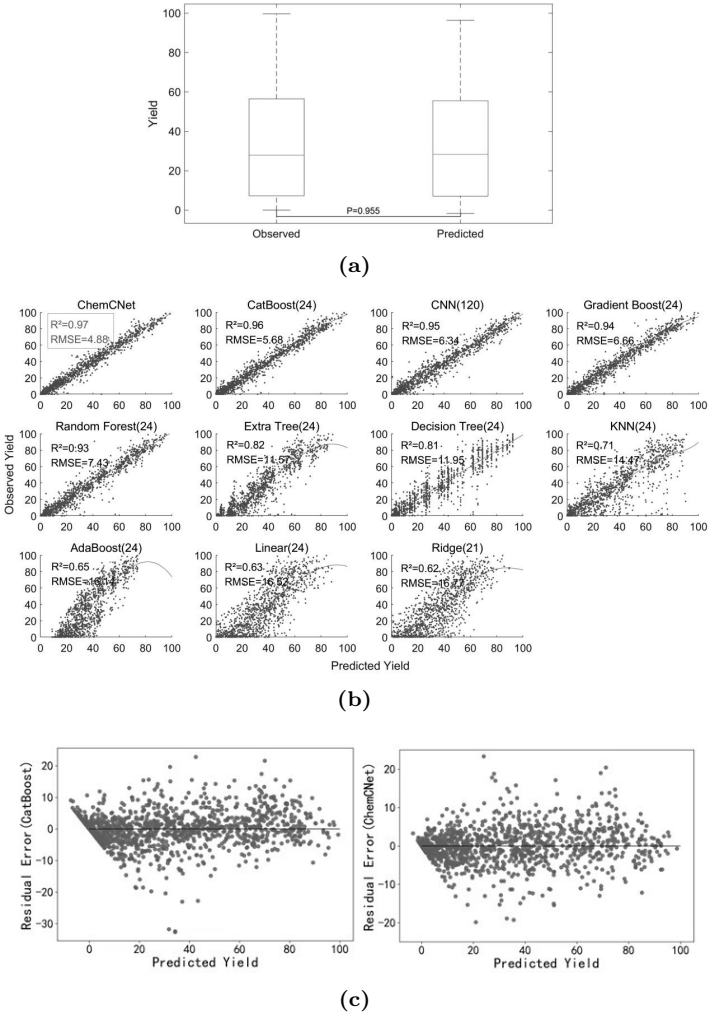
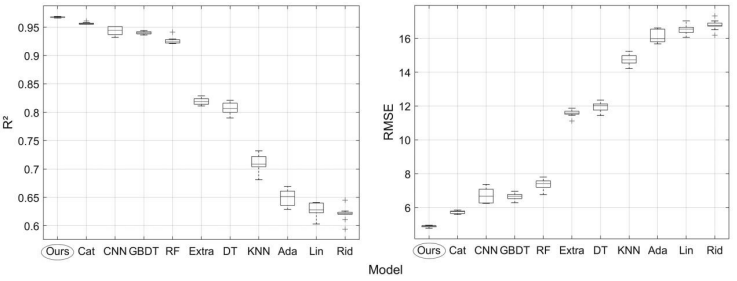
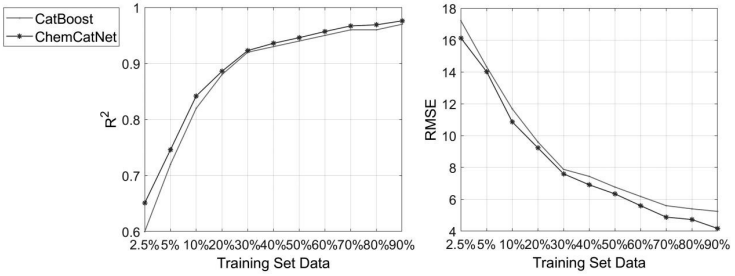


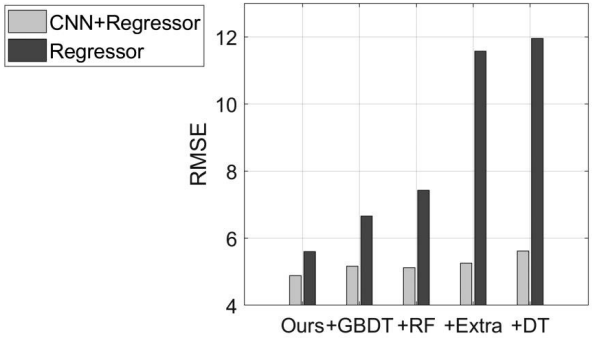

**Figure 5.** Prediction accuracy analysis. (a) Student's t test of the actual and predicted yield. (b) Comparison of predicted value and real value fitting scatter plot. (c) Comparison of regression residuals between ChemCNet model and CatBoost model.




**(a)**




**(b)**




**(c)**

**Figure 6.** Analysis of ChemCNet model prediction performance. (a) Box plot of prediction accuracy of different models. (b) Comparison of the prediction results of ChemCNet model and CatBoost model under different proportion of training data. (c) RMSE value comparison between ChemNet-0+machine learning regression and single machine learning regression.

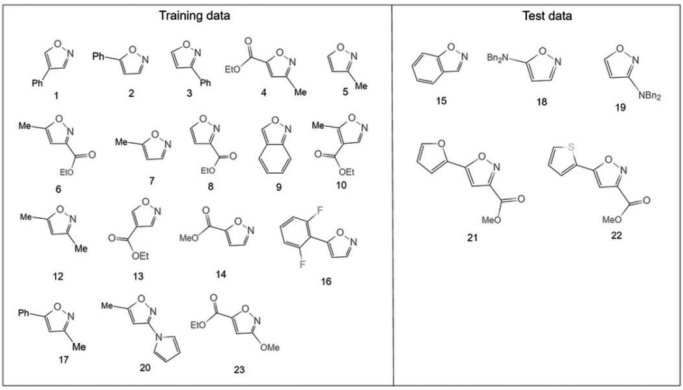
**Table 3.** Ablation experiment.

| | Feature Learning | CatBoost | Attention mechanism | $R^2$ | RMSE |
|---|---|---|---|---|---|
| 1 | | ✓ | | 0.961 | 5.68 |
| 2 | ✓ | | ✓ | 0.9668 | 4.93 |
| 3 | ✓ | ✓ | | 0.9671 | 4.91 |
| **4** | ✓ | ✓ | ✓ | **0.9674** | **4.88** |

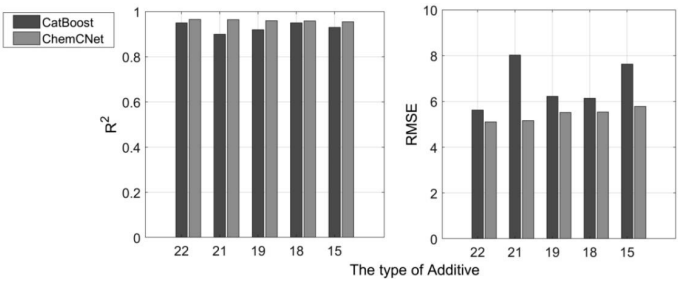
## 3.6 Generalization analysis

To evaluate the generalization performance of the ChemCNet, out-of-sample prediction experiments are conducted. 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. [7], five additives(15,18,19,21,22) are selected as unknown reaction conditions and the remaining known reaction conditions are 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 Fig.7.(a), and the out of sample prediction results are shown in Fig.7.(b).

As shown in Fig.7.(b), the out-of-sample predictions of the ChemCNet prediction model for all five additives are significantly higher than those of CatBoost, on average, no additive has significant systematic deviation from the prediction of the model. The high predictive power of ChemCNet indicates that the effects of these substituents on the reaction results can be well capture by the descriptors. In other words, the model constructed in this paper can predict the effect of new isoxazole or aryl halide structures on the outcome of Buchwald-Hartwig coupling reaction and provide the combination of bases and ligands with the highest yields.This also proves once again the excellent feature learning ability of deep learning and the effectiveness and significance of using it for feature learning. It also reflects from the side that features have a very important influence and role on model learning, the improvement of the overall performance of the model and the quality of the final experimental results.




**(a)**




**(b)**

**Figure 7.** Generalization performance analysis of ChemCNet. (a) Structure diagram of five out-of-sample predicted additives. (b) Out of sample prediction results of ChemCNet.

## 3.7 Interpretability analysis

In this chapter, we carry out three interpretability tools(feature importance, ALE, SHAP) based on ChemCNet model, aiming at providing comprehensive decision information for the experimenter. Since ChemCNet-0 obtains abstract features, which is not conducive to analysis, we use 24 feature descriptors obtained in Section 3.2 for interpretability analysis.

### 3.7.1 Feature importance analysis

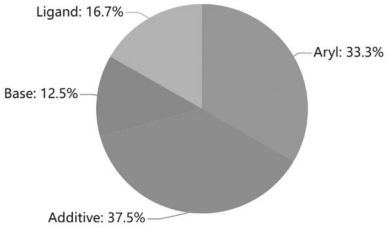
After obtaining the prediction model, we attempt to understand the factors that have a significant influence on the reaction yield prediction and

provide valuable information for improving the yield of Buchwald-Hartwig coupling reaction. Fig.8.(a) shows the class distribution of these 24 characteristic descriptors, and the results show that Additive (additive) and Halide (Aryl) account for the largest proportion, which may be the main factors affecting the yield prediction. Fig.8.(b) shows the feature importance ranking of CatBoost model. It can be seen that the top 10 most important descriptors for predicting reaction results include 5 aryl halides, 5 halides, 2 additives, 2 ligands and 1 substrate. Seven descriptors are related to electronegativity and NMR shift, including C-3 and H-2 electrostatic charges on halides, C-3 NMR shift on additives, C-4 electrostatic charges on additives, C-8 NMR shift and C-5 magnetic resonance shift on ligands, and N-1 electrostatic charges on substrates. The above analysis shows that the tendency of additives [24,25] and halides [26] as electrophiles may affect the reaction results, and the electronic effect of ligands also plays a key role in regulating the catalytic performance of metal catalysts.

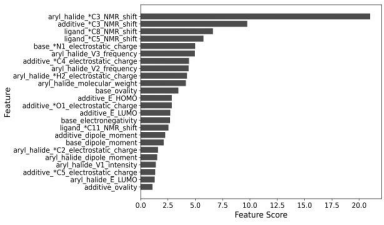
To validate the effectiveness of the features, we sampling 70% as training set, and the top 23-15 descriptors are selected, which are based on the feature ranking from high to low, as features to retrain CatBoost. The sampling is repeated for ten times generate ten results used to plot with corresponding feature numbers (noting the label as 0, right). 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 (notation labeled as 1,left) .

As can be seen from Fig.8.(c), as the number of features is decrease from 23 to 15, the prediction accuracy obtained by using the filter features for training remains at a high level, and the box volumes are all small, indicating that the distribution of the prediction results is more concentrated and stable, with no obvious fluctuations or differences. On the contrary, as the number of randomly selected features decreases, its prediction accuracy decreases significantly and the box volume increases, indicating that the distribution of its experimental results is more instable, with many outliers and less stable experimental results. In contrast, the radar plots (Fig.8.(d)) of prediction accuracy obtained using the features selected by feature importance for training are very regular, and the experimental re-

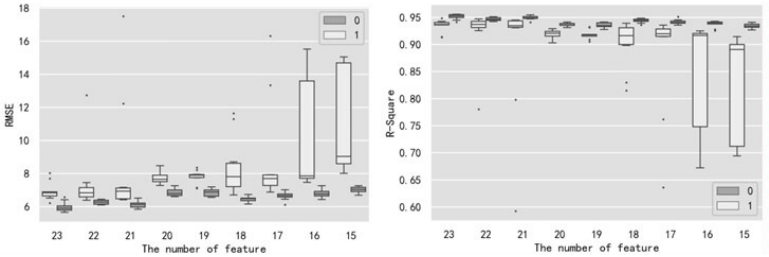
sults are all better than the randomly selected ones. In summary, the features cover comprehensive and important information, and the experimental results are more stable, which again verifies the effectiveness of the filtered features obtained.



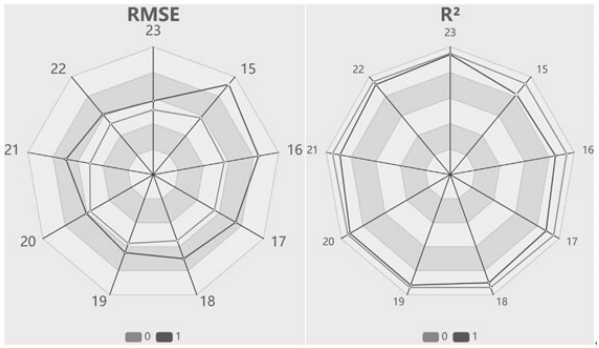

**(a)**




**(b)**




**(c)**




**(d)**

**Figure 8.** Feature importance analysis. (a) Category distribution of 24 descriptors. (b) Feature importance ranking. (c) Prediction accuracy boxplot. (d) Prediction accuracy radar plot.

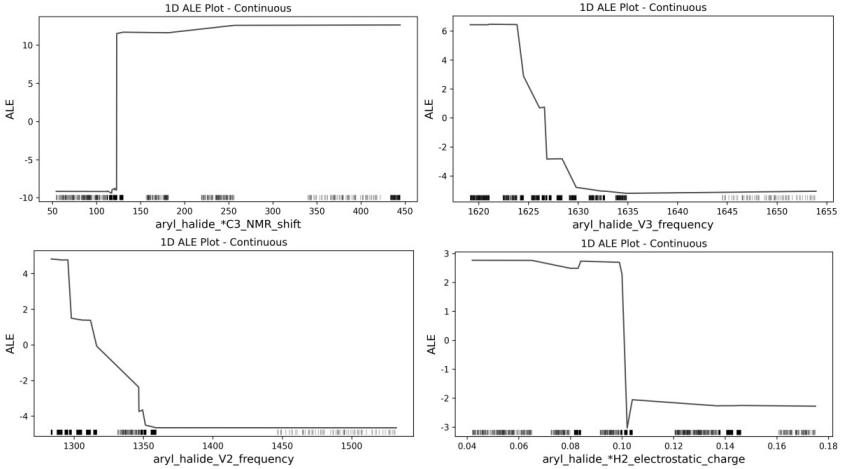
### 3.7.2 ALE-based analysis of the relationship between features and yield

Among all descriptors, halides account for the largest proportion and the descriptors representing the electronegativity of halides rank high, Then observe the relationship between aryl halides and reaction yield. Combined with the feature importance obtained in 3.7.1, we select feature descriptor aryl_halide_*C3_NMR_shift, aryl_halide_V3_frequency, aryl_halide_V2_frequency, aryl_halide_*H2_electrostatic_charge among the top 10 features for analysis.

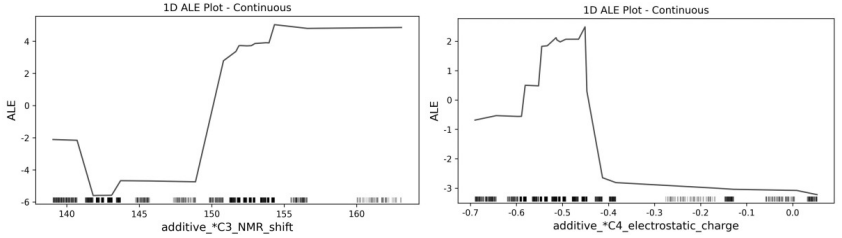
As shown in Fig.9.(a), (1) The ALE values of aryl_halide_*C3_NMR_shift and aryl_halide_*H2_electrostatic_charge changed more dramatically, both showing "linear changes". Among them, the ALE value of the descriptor aryl_halide_*C3_NMR_shift has the largest change amplitude, that is, the descriptor has the greatest influence on the prediction results, and when the feature value is taken between 100 and 150, the ALE value of the descriptor rises in a straight line from negative value to positive value, and then the ALE value remains in a high and stable state, indicating that the effect of the feature descriptor at this stage is a positive effect, and the reaction yield output by the model is higher than the average yield value. However, the descriptor has a linear decline when the eigenvalue is about 0.1. With 0.1 as the cut-off point, when the eigenvalue is less than 0.1, the reaction yield is 3 greater than the average predicted value, which plays a positive role, and when the eigenvalue is greater than 0.1, the reaction yield is lower than the average predicted value, which has a negative impact. (2) The ALE value of aryl_halide_V3_frequency and aryl_halide_V2_frequency has different amplitudes, so the degree of influence on yield is different, and the ALE value of the former changes more dramatically. The above analysis shows that electrophilic halides may have a strong effect on the reaction yield, and the vibration frequency of aryl halides in different vibration modes is not the same, and the effect on the reaction yield is also different.

Among all eight additive descriptors, four represent electronegativity properties, among which additive_* C3_ NMR_ shift and additive_* C4_ electrostatic_ Charge ranked second and sixth in importance respectively.

So the next step is to observe the relationship between additives and reaction yield. Same as above, select descriptor additive_* C3_ NMR_ shift, additive_* C4_ electrostatic_ charge for analysis. As can be seen from Fig.9.(b), the descriptor additive_* C3_ NMR_ Shift has the largest change in ALE value, that is, it has the largest impact on the yield. After the characteristic value is 150, the ALE value of this descriptor gradually increases, and after that, the ALE value remains at a high level, indicating that this descriptor has a positive impact on the reaction yield in the subsequent stage. The main reason for this difference is that additives, as an electron-rich system, strongly affect the reaction results.




**(a)**




**(b)**

**Figure 9.** ALE analysis. (a) ALE diagram of the influence of aryl halide descriptor on the reaction yield. (b) ALE diagram of the influence of additive descriptor on the reaction yield.

### 3.7.3 SHAP-based analysis of the relationship between features and yield

(1) Correlation analysis between features and yield

It is not enough to understand the importance of a feature, we do not know how the feature affects the prediction results, so it is necessary to further understand the correlation between the feature and the yield. SHAP profile maps analyze the correlation between feature descriptors and reaction yields. The overview diagram of the overall characteristics of SHAP is shown in Fig.10.(a), each row represents a feature descriptor with SHAP value, a dot represents a sample, and a wide area indicates that it contains a large number of samples. The color represents the feature value of the feature descriptor, and the shade of the color represents the feature value from small to large. It can be observed from the figure that the feature descriptor aryl_halide_*C3_NMR_shift has a small number of dark sample points on the right, but most of the light points are gather on the left, so there is basically a positive correlation between the feature descriptor and the reaction yield, that is, the larger the feature value of the descriptor, the larger the reaction yield. It is also noted that the characteristic descriptor ayrl_halide_V3_frequency also has a significant effect on the chemical reaction yield, and most of the light dots are concentrated on the right side, while only a few dark dots are concentrated on the left side, which means that there is a negative correlation between this descriptor and the reaction yield, that is, the larger the characteristic value of the descriptor, the smaller the corresponding reaction yield. The correlation between the other descriptors and the reaction yield is the same as above, so we will not analyze them one by one here.

(2) Feature interaction analysis

The combination of reaction conditions is very important in chemical reactions, and quantifying these interactions and revealing the hidden internal relationships is necessary and will help provide researchers with richer experimental information. The previous analysis shows that halide and additive may have a significant effect on the yield, especially aryl_halide_*C3_NMR_shift and additive_*C3_NMR_shift are important in the model prediction process, therefore, this subsection uses SHAP values

to further analyze the effect of the two features together on the reaction yield. The Shapely interaction index $\phi_{ij}$ extends the Shapely value by assigning credit among all feature pairs, on which the SHAP interaction value is defined as:

$$\Phi_{ij} = \sum_{S \subseteq M \setminus (i,j)} \frac{|S|!(|M| - |S| - 2)!}{2(|M| - 1)!} \nabla_{ij}(S), \tag{13}$$

where $\nabla_{ij}(S) = f(S \cup \{i,j\}) - f(S \cup \{i\}) - [f(S \cup \{j\} - f(S)], i \neq j$, $f(S) = E[f_x(x)|x_s]$.

Fig.10.(b) shows the SHAP dependency diagram of the two descriptors. In fact, the default second feature is automatically selected, that is, trying to pick out the feature column that interacts most strongly with the additive_*C3_NMR_shift. When the second feature is not specified, the figure automatically selects the aryl_halide_*C3_NMR_shift, indicating that there is indeed a strong interaction between the two descriptors. The X axis is the eigenvalue range of the feature descriptor additive_*C3_NMR_shift, the Y axis is its SHAP value, and the right is the eigenvalue range of the contrasting descriptor aryl_halide_*C3_NMR_shift, where dark represents the high eigenvalue part of the descriptor and light represents its low eigenvalue part. Color analysis is the distribution of aryl_halide_*C3_NMR_shift in the process of additive_*C3_NMR_shift changes. The light gray area at the bottom of the plot shows a histogram of the distribution of data values.

It can be observed from the figure that the values of the feature points of the descriptor additive_*C3_NMR_shift are concentrated between the intervals (140,145) and (150,155). When the descriptor additive_*C3_NMR_shift takes a value before b, there are more light dots and fewer dark dots, and most of its SHAP values are negative, indicating that for the descriptor additive_*C3_NMR_shift before b, the smaller the value of the descriptor aryl_halide_*C3_NMR_shift, the descriptor additive_*C3_NMR_ The greater the negative impact of shift on reaction yield.

When the descriptor additive_*C3_NMR_shift takes values between $(b, 150)$, the SHAP value of the descriptor additive_*C3_NMR_shift is negative, regardless of how the descriptor aryl_halide_*C3_NMR_shift is taken, that is, it negatively affects the reaction yield. At the same time, it was

noted that the dark dots were distributed below, which indicated that the higher the eigenvalues of the descriptor aryl_halide_*C3_NMR_shift, the more likely it was to have a negative impact on the reaction yield.

When the descriptor additive_*C3_NMR_shift is taken after 150, the SHAP value of the descriptor additive_*C3_NMR_shift is positive regardless of the value of the feature descriptor aryl_halide_*C3_NMR_shift, that is, it has a positive effect on the reaction yield. And the dark dots are mostly distributed above at this time, which indicates that the higher eigenvalues of the descriptor aryl_halide_*C3_NMR_shift are more likely to have a positive impact on the reaction yield.

From the above analysis, it can be concluded that when the descriptor additive_*C3_NMR_shift takes a value after 150, the higher the aryl_halide_*C3_NMR_shift, the greater the positive effect on the yield, and the easier it is to obtain a high yield. Combined with the previous correlation analysis between the characteristic variables and the reaction yield, it can be seen that these two feature descriptors are positively correlated with the reaction yield. However, from the interaction analysis, it is known that even "positive cooperation" is limited by "specific conditions" (i.e. different values) to obtain the ideal reaction yield.

After that, we show the reaction yield obtained under different combinations of these two feature descriptors. The results are shown in Fig.12, the horizontal axis is the value range of the descriptor additive_*C3_NMR_shift, the vertical axis is the value range of aryl_halide_*C3_NMR_shift, and the more yellow the color of the point, the higher the reaction yield obtained. We are more concerned about high yield than low yield, and it can be observed from the figure that when the value of additive_*C3_NMR_shift is 153.74 and the value of aryl_halide_*C3_NMR_shift is 256.46, the reaction yield reaches a maximum of 99.03. Through the analysis of ALE values in the previous section, it can also be found that these two values correspond to the situation that their respective ALE values are greater than 0, which is consistent with the conclusions obtained from the previous analysis.

(3) SHAP analysis of one single sample

Because of the variability among individuals, it is necessary to un-

derstand the feature effects exhibited by features in the sample prediction process. The force plot provides the details of the prediction, and it focuses on explaining how individual predictions are generated and how individual features affect the model's decision in a single instance. The longer the arrow, the greater the impact of the feature on the output. Base_value is the average predicted value of all samples, and output_value, or $f(x)$, is the predicted value of the sample. A sample is randomly selected for illustration, as shown in Fig.10.(c), aryl_ halide_* C3_ NMR_ Shift and base_* N1_ electrostatic_ Charge is the two descriptors with the largest positive contribution to the predicted value; aryl_ halide_ V3_ Frequency and aryl_ halide_ V2_ Frequency is the two descriptors with the largest negative contribution to the predicted value. However, it is noted that when there are many features, the feature effect of each feature cannot be fully displayed. Therefore, we consider using waterfall diagram for visual display.

The waterfall plot powerfully shows how a sample accumulates from the base_value to the final prediction of the model at the top of the plot, while giving the magnitude and direction of the influence of each feature. Fig.11.(a) shows the waterfall plot for this sample, which shows the prediction process of this sample and the respective contribution of each feature in the model prediction process. $f(x)$ indicates the final predicted value of this sample. The values next to the descriptor names are their feature values. Starting from the base value of 32.94 at the bottom of the waterfall chart, the dark SHAP values indicate an increase in prediction and the light SHAP values indicate a decrease in prediction. Compare to the force diagram, the waterfall diagram shows more concisely the features that play an important role in the model prediction process and their "contribution".

This will provide researchers with a more comprehensive and detailed understanding of the specific effects of the descriptors in a specific chemical reaction, so that they can make adjustments to their experiments.

(4) Multisample clustering SHAP analysis

The purpose of clustering is to find samples with certain similarities. Generally speaking, clustering is based on features. This subsection uses a stacked SHAP force map for clustering to analyze the influence of different

features of many samples. The figure is obtained by rotating the force map of a single sample by 90 degrees and stacking it horizontally. Due to the large sample size, it is not easy to show all of them, so only the predictions of the first 100 samples are shown here. The result is shown in Fig.10.(b), the vertical axis is the predicted value $f(x)$, the horizontal axis indicates the number of samples, and the horizontal axis is aggregated and arranged by the similarity of the samples. The graph will aggregate the samples with similar SHAP values, thus the horizontal coordinate indicates the serial number of the sample rearrangement, and the original index of this sample will be displayed after mouse click. Each sample consists of a dark area and a light area, with the light shaded part indicating the negative impact on the model output and the dark shaded part indicating the positive impact on the model output. The larger the dark area, the stronger the positive impact, and conversely, the larger the light area, the stronger the negative impact. The different features and their SHAP values can be seen by pointing the mouse to any position at random. The clustering is highlighted: the center part of the figure (serial number between 30 and 50) has more light area, which is the output of the model, that is, the response with lower than average response yield; the second half of the figure (serial number between 60 and 80) has more dark area, which is the response with higher than average response yield. In Fig.11.(a), the SHAP value of the 71st sample point (actually indexed as 13) is shown as an example, and the final predicted value of this sample is 79.59 higher than the average predicted value under the combined effect of each feature.

In practice, experimenters can use the graph to cluster samples of reactions with similar properties, thus eliminating the need to view the reaction of each chemical reaction one by one. By looking at their common characteristics, the adjustment can be effectively narrowed down and scientific efficiency can be improved, while saving experimental resources.

In the Buchwald-Hartwig coupling reaction, the electrophilic reagents are able to undergo oxidative addition reactions with the zero-valent palladium complexes to produce transition state compounds of divalent palladium for the final production of the products. Among them, additives and halides as electrophilic reagents are two important components affect-

ing the yield of the reaction, which is consistent with the objective facts and indicates that the output of the model constructed in this paper is real and creditable. It is worth reminding that it is necessary to combine the conclusions drawn from the above analysis with the experience of the researchers and some specific experimental settings for comprehensive decision making.

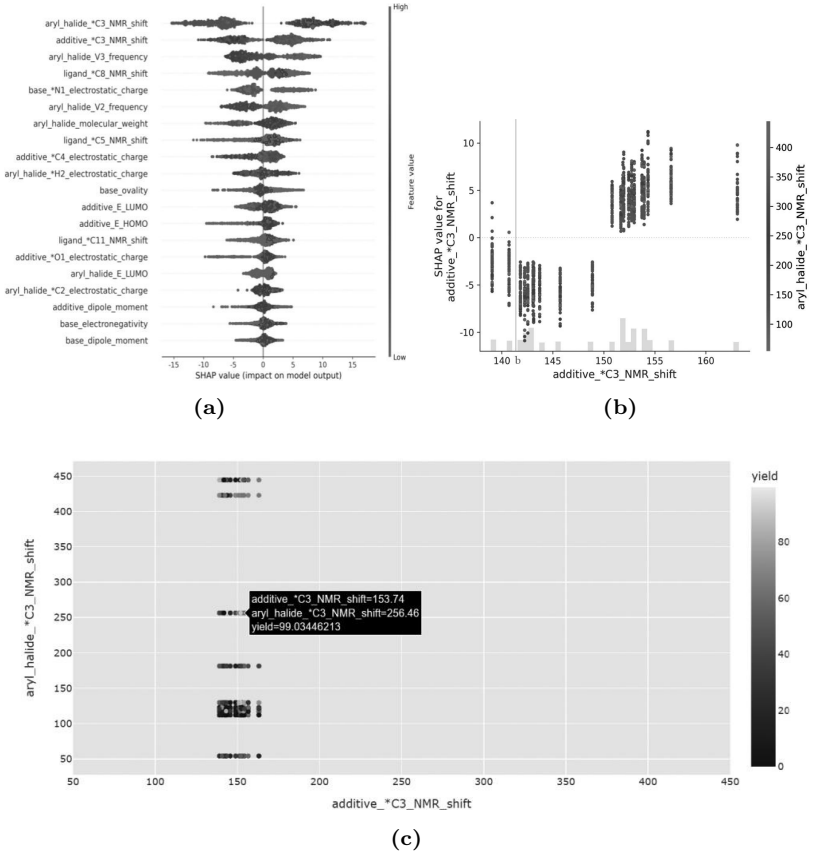
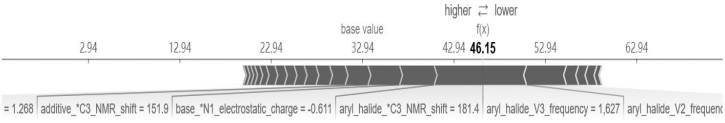
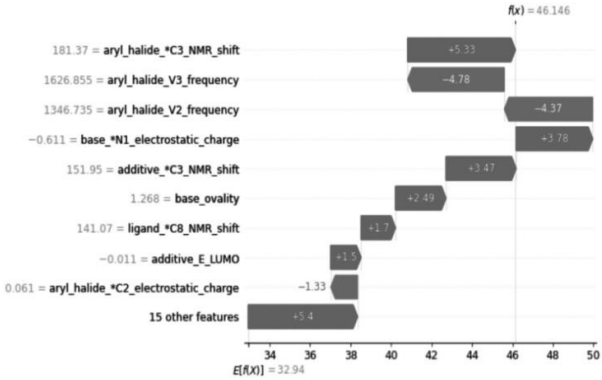



**Figure 10.** SHAP analysis. (a)SHAP summary plot. (b) SHAP dependency graph of two feature descriptors. (c) The reaction yield under different value combinations of feature descriptors.




**(a)**




**(b)**

**Figure 11.** SHAP analysis of one single sample. (a) SHAP force plot. (b) SHAP waterfall.

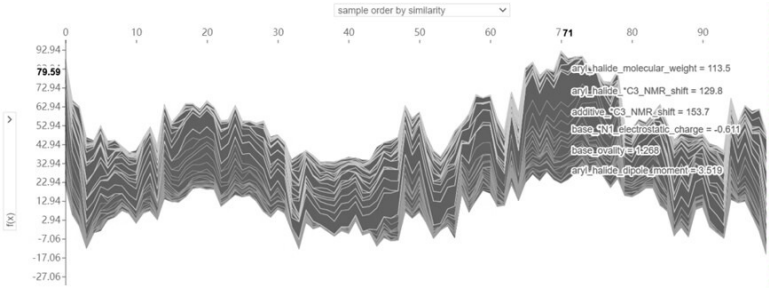



**Figure 12.** SHAP cluster analysis.

As shown in Fig.13, for the convenience of users, we has developed a free EXE software called ChemCNet, which can implement feature analysis, reaction yield intelligent prediction and interpretability analysis.

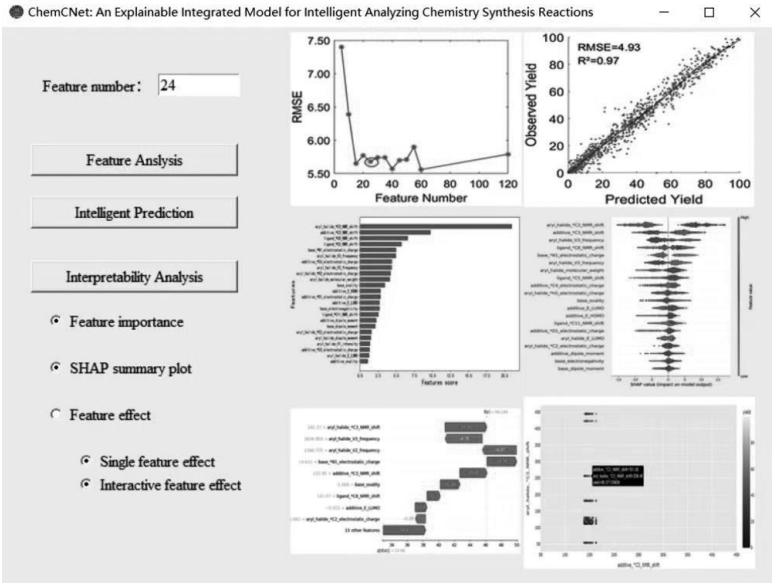



**Figure 13.** The system interface of ChemCNet.

# 4 Conclusions

In this paper, a model-related feature selection method is adopted, and 24 feature descriptors with a large impact on the reaction yield are screened out. Based on this, combine CNN with CatBoost, adding attention mechanism, and finally build ChemCNet prediction model. ChemCNet enhances the expression ability of features through feature re-representation, and it provides more decision information for the experimenter with the good interpretability of CatBoost . The experiments show that ChemCNet prediction model achieves high accuracy in predicting reaction yields and has good generalization ability. This will better assist the progress of research in chemistry disciplines and provide more accurate help to experimenters.

# References


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

[2] E. Negishi, A. O. King, N. Okukado, Selective carbon-carbon bond formation via transition metal catalysis. 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.

[3] A. O. King, N. Okukado, E. I. 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.

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

[5] 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.

[6] V. Venkatasubramanian, V. Mann, Artificial intelligence in reaction prediction and chemical synthesis, *Curr. Opin. Chem. Eng.* **36** (2022) #100749.

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

[8] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, Generating focused molecule libraries for drug discovery with recurrent neural networks, *ACS Cent. Sci.* **4** (2018) 120–131.

[9] 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.

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

[11] X. H. Yang, W. M. Wu, Y. M. Chen, X. Q. Li, J. Zhang, D. Long, L. J. Yang, An integrated inverse space sparse representation framework for tumor classification, *Pattern Recogn.* **93** (2019) 293–311.

[12] M. Chieregato, F. Frangiamore, M. Morassi, C. Baresi, S. Nici, C. Bassetti, C. Bnà , M. Galelli, A hybrid machine learning/deep learning COVID-19 severity predictive model from CT images and clinical data, *Sci. Rep.* **12** (2022) 1–15.

[13] C. W. Coley, W. Jin, L. Rogers, T. Jamison, T. Jaakkola, W. Green, R. Barzilay, K. Jensen, A graph-convolutional neural network model for the prediction of chemical reactivity, *Chem. Sci.* **10** (2019) 370–377.

[14] H. X. Hou, H. Z. Wang, Y. H. Guo, P. Y. Zhang, L. C. Peng, X. H. Yang, Regression prediction of coupling reaction yield based on attention–driven convolutional neural network, *MATCH Commun. Math. Comput. Chem.* **89** (2023) 199–222.

[15] Y. N. Zhao, X. C. Liu, H. Lu, X. F. Zhu, T. H. Wang, G. Luo, R. C. Zheng, Y. Luo. An optimized deep convolutional neural network for yield prediction of Buchwald-Hartwig amination, *Chem. Phys.* **550** (2021) #111296.

[16] Z. T. Song, D.Y. Huang, B. W. Song, K.Q. Chen, Y. Y. Song, G. Liu, J. L. Su, João Pedro de Magalhães, D. J . Rigden, J. Meng, Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications, *Nat. Commun.* **12** (2021) #4011.

[17] D. W. Apley, J. Zhu, Visualizing the effects of predictor variables in black box supervised learning models, arXiv:1612.08468, 2016.

[18] S. M. Lundberg, S. I. Lee. A unified approach to interpreting model predictions, *NIPS.* **31** (2017) 4768–4777.

[19] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, *Mach. Learn.* **46** (2002) 389–422.

[20] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, *Comput. Sci.* **31** (2018) 6637–6647.

[21] A. V. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support, arXiv:1810.11363, 2018.

[22] Q. L. Wang, B. G. Wu, P. F. Zhu, P. H. Li, W. M. Zuo, Q. H. Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks, *IEEE Conf. Comput. Vis. Pattern Recogn.* (2020) 11534–11542.

[23] T. Q. Chen, G. Carlos, XGBoost: a scalable tree boosting system, *ACM* **22** (2016) 785–794.

[24] 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.

[25] M. Shigenobu, K. Takenaka, H. Sasai, Palladium-catalyzed direct C–H arylation of isoxazoles at the 5-position, *Angew. Chem. Int. Edit.* **54** (2015) 9572–9576.

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