Machine Learning of C-H Activation Reaction of Indoles

Chao Liu[1], Yifei Wang[2]

1.Department of Chemistry, Southern University of Science and Technology of China, Guangdong, Shenzhen, 518055, China
2.Shenzhen College of International Education, Guangdong, Shenzhen, 51800, China

ABSTRACT. *Aiming at developing the most suitable method based on machine learning to predict the results of C-H activation reaction of indoles (CARI), we constructed the database by using ortho-substitution and meta-substitution to collect 2000 different CARI. DFT method is employed to calculate energy barrier and selectivity of all reactions.7 algorithms are involved -Random forest (RF), Kernel Ridge Regression (KRR), K Nearest Neighbor (KNN), Support Vector Regression (SVR), Neural Network (NN), Gaussian Process (GP) and LASSO Regression (LASSO)-to make predictions. Meanwhile, the results of the reactions give us the guidance to predict the energy barrier and selectivity of new reactions for experimentalists. According to the data, Random forest is the most suitable way to make predictions. Superior to random forest, KNN and NN also demonstrates good performance.*

KEYWORDS: *Machine learning, C-H activation, Indoles, Ortho-substitution, Meta-substitution*

1. Introduction

The C-H bond is a very common type of chemical bond. Almost all organic compounds contain C-H bonds. The bond energy of the CH bond is very high, and the electronegativity of carbon and hydrogen is very close, so the polarity of the CH bond is very small. These factors make it inert. The CH bond can be selectively activated and constructed under mild conditions. Other carbon-containing chemical bonds have dual challenges of thermodynamics and kinetics, which are a basic problem in chemistry research and a bottleneck problem that restricts major breakthroughs in molecular synthesis and preparation. In recent years, with the rapid development of transition metal-catalyzed C-H bond activation, the research on the construction of indole rings through C-H bond construction has aroused widespread interest among chemists.

Indole (C8H7N) is a compound of Pyrrole (C4H5N) and Benzene (C6H6) in parallel. This compound of pyrrole and benzene in parallel is also called benzo pyrrole. There are two ways to combine pyrrole and benzene in parallel, known as indole and isoindole respectively. The indole structure is widely present in natural products and biologically active molecules. It is also important in organic synthesis intermediate. Due to the significant physiological activity of the indole structure, it has become one of the important precursors for new drug design and development. Some compounds have been developed into clinical drugs. In terms of drug synthesis, as a structural skeleton of a chemical drug, indole, with its high biological activity, makes it useful in anti-hypertensive, anti-proliferative, anti-viral, anti-Tumor, Analgesie, anti-inflammatory, anti-bacterial, etc. All the drugs in the therapeutic field have a place, and organic chemists are therefore very interested in it.

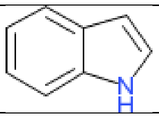
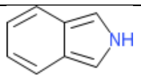
	
1-a)Indole structural formula	1-b)Isoindole structural formula

Fig.1 Two Combinations of Pyrrole and Benzene

When the world's first general-purpose computer "ENIAC" was born on February 14, 1946, people officially entered the electronic age. In 1956, McCarthy first proposed the concept of artificial intelligence (AI), which opened up this new field of research. For more than half a century, as the computing skill of computers has improved dramatically, machine learning has developed rapidly and is widely used in various fields. With the development of cheminformatics, machine learning has shown great potential for the applications in the field of chemistry. The field of organic chemistry has achieved gratifying breakthroughs in the synthesis of complex macromolecules through the application of machine learning.

Due to the special nature of the mentioned C-H bond, the energy barrier and selectivity of the C-H bond are difficult to calculate. The results need to be gradually obtained through continuous experiments, which bring unnecessary trouble to the experimental process.

Based on the characteristics of machine learning, the energy barrier and selectivity of indoles c-H bond activation reaction can be calculated by machine learning before the experiment, which provides great convenience for the subsequent research work.

2. Research Methods

Machine learning is a multidisciplinary subject. The machine learning of the indole C-H activation reaction combines organic chemistry and computer science, using machine learning to predict the energy barrier and selectivity of the indole C-H bond activation reaction. After searching the Internet and local literature and using actual experimental results, a total of 2000 detailed data on the energy barrier and selectivity of the indole CH bond activation reaction were collected. We than use the collected indole C-H bond activation reaction energy barrier and selectivity data to build the energy barrier and selectivity database of indole C-H bond activation reaction for machine learning.

The experimental ideas on the energy barrier and selectivity of machine learning indoles C-H bond activation reaction are as follows: Use different kinds of machine learning algorithms to calculate the energy barrier and selectivity of each indole CH bond activation reaction. Classify the energy barrier and selectivity of 2000 indole CH bond activation reactions according to meta substitution and ortho substitution. The machine learning model is established through a variety of algorithm training data to predict the energy barrier and selectivity of the new indole CH bond activation reaction and find the most suitable calculation method for the indole CH bond activation reaction result based on machine learning.

In the selection of substituent groups, a total of 10 types of substituents in four major applications were selected to participate in the reaction. The specific substituents are as follows:

Table 1 Ten Substituents in Experiment

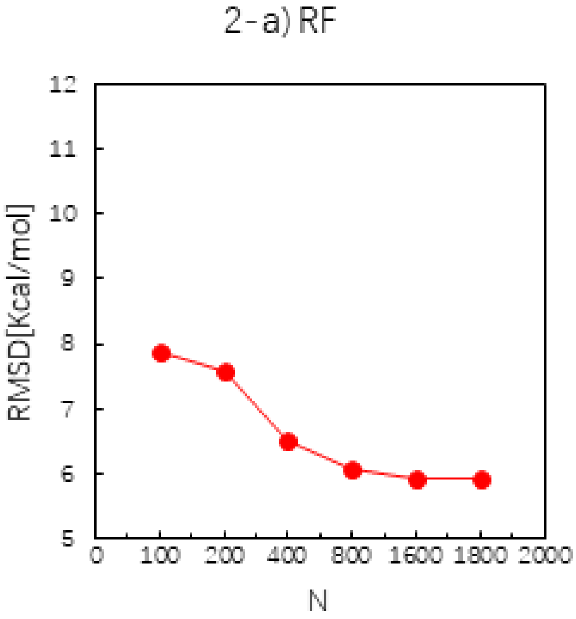
Electron-donating group	-NH_2	(Amino)	Resistance group	-CH_3	Methyl
	-NME_2	N,N-Dimethyl-2-(methylamino)Acetamide		-$C(CH_3)_3$	tertiary butyl
	-OH	Hydroxy	Strong electron withdrawing group	-$COCH_3$	acetyl
	-OCH_3	Methoxyl group		-NO_2	nitroso
Electron-withdrawing group	-F	Fluorine			
	-Br	Bromine			

Since the core of AI is algorithms, machine learning has been developed into a wide variety of algorithms. After screening and comparing dozens of machine learning algorithms, random forest (RF) [5], Kernel Ridge Regression (KRR), and K Nearest Neighbor (KNN) Support Vector Regression (SVR), Neural Network (NN), Gaussian Process (Gaussian Process, GP) and LASSO regression (LASSO) 7 kinds of algorithms are selected, which are suitable for indole C-H bond activation reaction energy barrier and selectivity calculation requirements. We calculate the energy barrier and selectivity of each reaction.

3. Experimental Data and Results

In machine learning, we use Root Mean Square Error (RMSD) to screen 7 algorithms that are most suitable for calculating the results of indole CH activation reaction based on machine learning. The Support Vector Regression (SVR) algorithm has been tested for several time and we find out that It is not suitable for machine learning modeling of indole CH activation reaction system, so this data is not presented in the algorithm learning curve.

The specific experimental data are as follows:

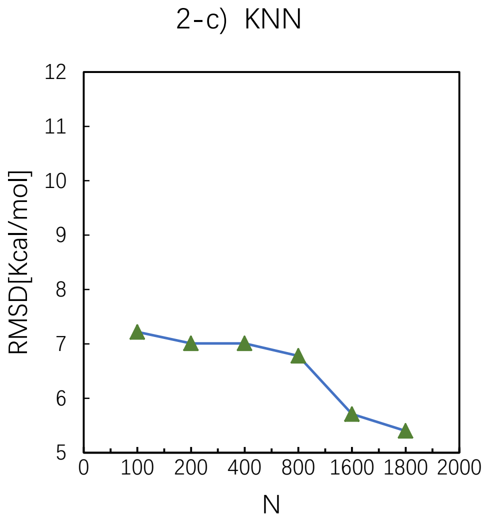


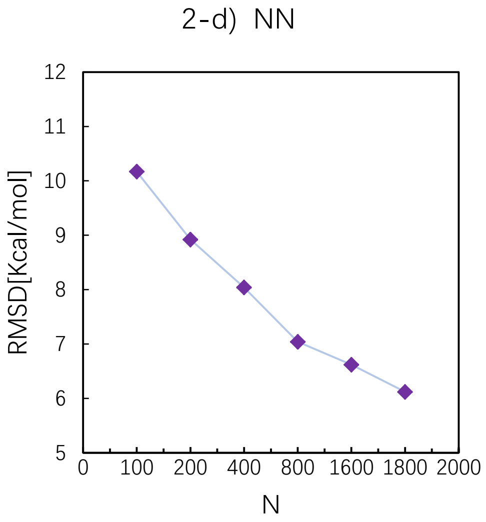
A.Random forest(RF):

N=100,RMSD=7.88Kcal/mol;N=200,RMSD=7.57Kcal/mol;N=400,RMSD=6.49 Kcal/mol;N=800,RMSD=6.07Kcal/mol;N=1600,RMSD=5.91Kcal/mol;N=1800,RMSD=5.90Kcal/mol.

B.Kernel Ridge Regression (KRR):

N=100,RMSD=11.21Kcal/mol;N=200,RMSD=7.63Kcal/mol;N=400,RMSD=7.44Kcal/mol;N=800,RMSD=6.96Kcal/mol;N=1600,RMSD=6.32Kcal/mol;N=1800,RMSD=6.17Kcal/mol.





C.K Nearest Neighbor(KNN):

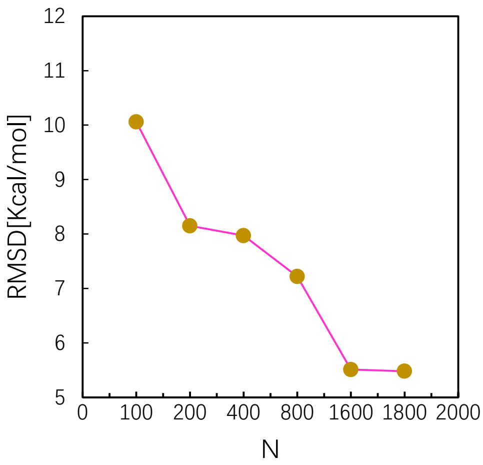
N=100,RMSD=7.22Kcal/mol;N=200,RMSD=7.01Kcal/mol;N=400,RMSD=7.01Kcal/mol;N=800,RMSD=6.78Kcal/mol;N=1600,RMSD=5.71Kcal/mol;N=1800,RMSD=5.40Kcal/mol.

D.Neural Network(NN):

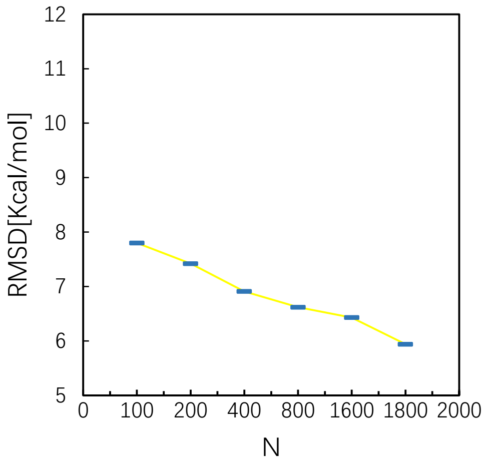
N=100,RMSD=10.17Kcal/mol;N=200,RMSD=8.92Kcal/mol;N=400,RMSD=8.04Kcal/mol;N=800,RMSD=7.04Kcal/mol;N=1600,RMSD=6.62Kcal/mol;N=1800,R

MSD=6.12Kcal/mol.

2-e) GP



2-f) LASSO



E.Gaussian Process (GP):

N=100,RMSD=10.06Kcal/mol;N=200,RMSD=8.15Kcal/mol;N=400,RMSD=7.97Kcal/mol;N=800,RMSD=7.22Kcal/mol;N=1600,RMSD=5.51Kcal/mol;N=1800,RMSD=5.48Kcal/mol.

F.LASSO Regrassion (LASSO):

N=100,RMSD=7.80Kcal/mol;N=200,RMSD=7.42Kcal/mol;N=400,RMSD=6.91 Kcal/mol;N=800,RMSD=6.62Kcal/mol;N=1600,RMSD=6.43Kcal/mol;N=1800,RMSD=5.94Kcal/mol.

Based on the above data, the learning curve of the six algorithms as shown in Figure 2 is drawn:

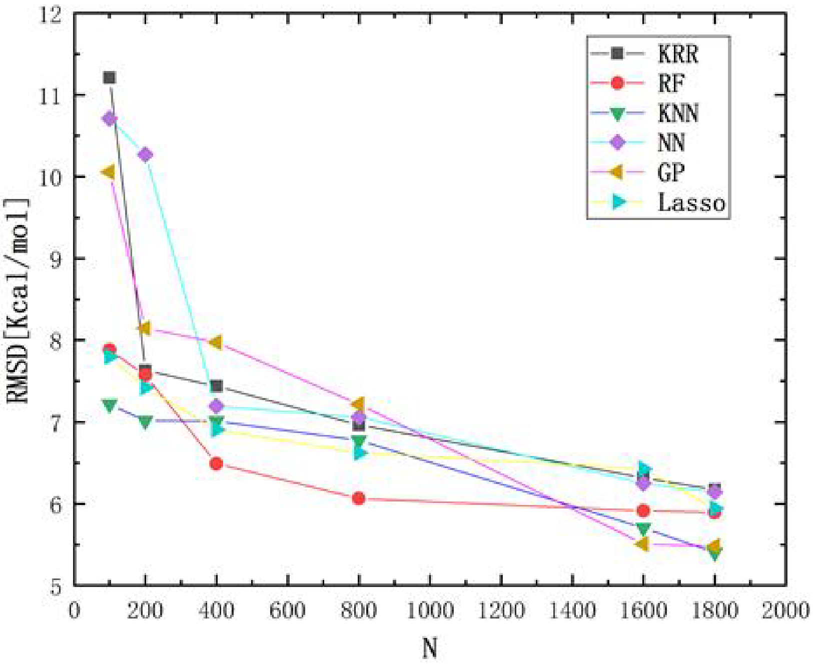


Fig.2 Learning Curve of the Six Algorithms

Among them, N represents the number of training sets, RMSD[Kcal/mol] represents the square root of the ratio of the square of the deviation between the predicted value and the true value to the number of observations n. As the training set N increases, the root mean square error can be observed. The smaller the root mean square error, the more accurate the result. It can be observed and judged from Figure 2 that random forest (RF) performs well for low training sets, and as the number of data sets increases, the performance of the learning curve of the six algorithms is smoother and the data is relatively stable. So then use the random forest (RF) algorithm to conduct targeted experiments.

Selective prediction of indoles C-H activation reaction was performed based on the Random Forest (RF) algorithm, and the results are shown in the figure below:

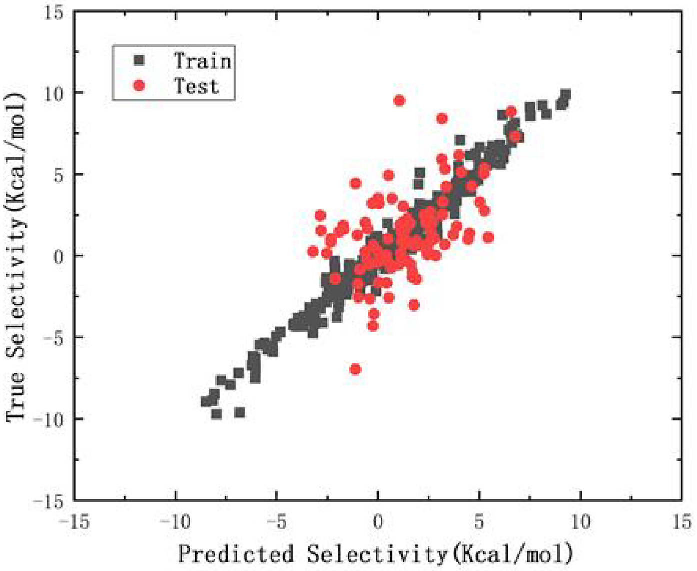


Fig.3 Selective Prediction Results of Random Forest

Scatterplot in Figure 3 shows the predicted selectivity and its corresponding true selectivity of Random Forest. The coefficient of determination, which is measured in [0 1] interval shows the explanation ability of models to dependent variable y. The coefficient of determination, which measured by R-squared, is 0.97 for training samples. It implies that Random Forest provide a well predicted result. In addition, the coefficient of determination for testing sample is 0.82, which is higher than 0.80. This result is also good and acceptable.

Meanwhile, the indoles C-H activation reaction barrier was predicted based on the Random Forest (RF) algorithm, and the results are shown in the figure below:

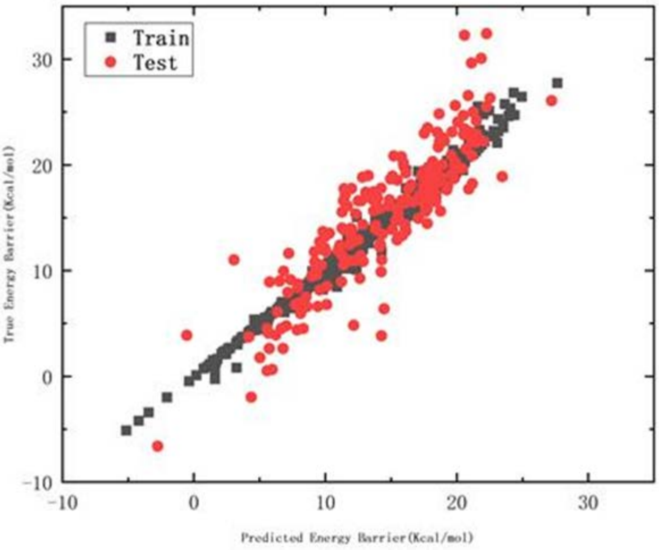


Fig.4 Predicted Energy Barrier Using Rf

The coefficient of determination is 0.99 for training samples. It implies that Random Forest provides a well predicted result for Energy Barrier too. Moreover, the coefficient of determination for testing sample is 0.85, which is higher than 0.80. This result is also good as well.

4. Conclusions

The experiments have obtained good experimental results. According to the above experimental data, it can be concluded that the random forest algorithm is the best method for the indole CH bond activation reaction energy barrier and selective machine learning. The random forest algorithm is used for machine learning. The results are stable and accurate, showing low fluctuation characteristics in the number of training sets given in the experiment. Using the random forest algorithm provides guidance for predicting the energy barrier and selectivity of the new indole C-H bond activation reaction, bringing convenience to the experimental process, laying a foundation for subsequent experiments, and further improving the accuracy of the experiment.

References

[1] Junfei Lu, Xing Xu, Junliang Zheng.Research progress of indole C - H arylation reaction [J].Organic Chemistry,2018(02):363-377.

[2] Beili Lu, Xianyan Li, Yongmei, Lin, Research progress in the formation of indole structure by c-H bond group catalyzed by transition metals [J].Organic Chemistry,2015,35(011):2275-2290.

[3] Lalit K, Shashi B, Kamal J. The diverse pharmacological importance of indole derivatives: a review[J]. International Journal of Research in Pharmacy & Science, 2012, 2(2): 22-33.

[4] Yidi Liu, Qi Yang, Yao Li.Application of machine learning in organic chemistry [J].Organic Chemistry,2020(0823):1-16.

[5] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5 - 32.