





Article

Machine Learning and Quantum Calculation for Predicting Yield in Cu-Catalyzed P–H Reactions

 Youfu Ma ¹, Xianwei Zhang ¹, Lin Zhu ¹, Xiaowei Feng ^{1,2}, Jamal A. H. Kowah ^{1,2} , Jun Jiang ¹ , Lisheng Wang ^{1,*}, Lihe Jiang ^{2,*}  and Xu Liu ^{1,2,*} 

¹ Medical College, Guangxi University, Nanning 530004, China; muduo_youfu@163.com (Y.M.); 13507804424@163.com (X.Z.); zll17851182520@163.com (L.Z.); fengxiaowei77@outlook.com (X.F.); jamal.kowah@gmail.com (J.A.H.K.); jiangjun@gxu.edu.cn (J.J.)

² School of Basic Medical Sciences, Youjiang Medical University for Nationalities, Baise 533000, China

* Correspondence: w_lsheng@163.com (L.W.); jianglihe@ymun.edu.cn (L.J.); wendaoliuxu@163.com (X.L.); Tel.: +86-13737071312 (L.W.); +86-18577190501 (L.J.); +86-15807710318 (X.L.)

Abstract: The paper discussed the use of machine learning (ML) and quantum chemistry calculations to predict the transition state and yield of copper-catalyzed P–H insertion reactions. By analyzing a dataset of 120 experimental data points, the transition state was determined using density functional theory (DFT). ML algorithms were then applied to analyze 16 descriptors derived from the quantum chemical transition state to predict the product yield. Among the algorithms studied, the Support Vector Machine (SVM) achieved the highest prediction accuracy of 97%, with over 80% correlation in Leave-One-Out Cross-Validation (LOOCV). Sensitivity analysis was performed on each descriptor, and a comprehensive investigation of the reaction mechanism was conducted to better understand the transition state characteristics. Finally, the ML model was used to predict reaction plans for experimental design, demonstrating strong predictive performance in subsequent experimental validation.

Keywords: machine learning; quantum chemistry; SVM; transition state; copper catalysts



Citation: Ma, Y.; Zhang, X.; Zhu, L.; Feng, X.; Kowah, J.A.H.; Jiang, J.; Wang, L.; Jiang, L.; Liu, X. Machine Learning and Quantum Calculation for Predicting Yield in Cu-Catalyzed P–H Reactions. *Molecules* **2023**, *28*, 5995. <https://doi.org/10.3390/molecules28165995>

Academic Editor: Dingyi Wang

Received: 1 July 2023

Revised: 30 July 2023

Accepted: 1 August 2023

Published: 10 August 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Catalysts play an essential role in various chemical transformations. However, the search for highly efficient catalysts for specific reactions remains a challenging task due to the complexity of catalytic processes [1–3]. One effective strategy for constructing C–C and C–heteroatom bonds is the insertion reaction of α -azido carbonyl compounds catalyzed by transition metals [4–8]. Another significant method for synthesizing organic phosphine compounds is the P–H insertion reaction [9]. However, there is relatively little research on P–H insertion reactions in comparison to other X–H insertion reactions, and the range of applicable metal catalysts is limited. The weak nucleophilic ability of phosphorus, along with its high susceptibility to coordination bonding with metals due to the presence of lone pair electrons in its outermost shell, poses challenges in the formation of metal carbene intermediates and the occurrence of P–H insertion reactions. Copper has emerged as a crucial catalyst in facilitating P–H insertion reactions [10–13]. Nonetheless, the complexity of the reaction, involving numerous catalysts and substrates, requires scientists to rely on their expertise and intuition, conducting trial and error experiments to identify suitable reaction conditions. Despite significant efforts and resources invested, the outcomes often prove unsatisfactory.

In the field of catalyst design and optimization, computational chemistry has become increasingly important [14]. One strategy involves using quantum chemical methods to simulate reaction transition states [15–23]. However, the vast space of catalytic materials and the diversity of reaction conditions make traditional quantum mechanical-based computational chemistry inefficient for catalyst screening [24–26]. Fortunately, artificial

intelligence (AI) technology based on machine learning algorithms can overcome these barriers, significantly accelerating the catalyst design process [27–29]. Integrating quantum chemistry transition-state models with machine learning in catalyst design workflows can provide valuable information, including experimental yield predictions and transition-state characteristics that may not be easily obtained through other means.

While there has been growing interest in using machine learning and quantum chemistry calculations to parameterize experimental data and predict optimal catalytic conditions [29–34], examples of using machine learning to analyze experimental data and predict results under new reaction conditions are still limited. The combination of these two approaches to predict product yields in copper-catalyzed P–H insertion reactions is a novel application.

In this study, we employ quantum chemical calculations to elucidate the transition state of the copper-catalyzed P–H insertion reaction (Figure 1). Subsequently, we integrate the quantum chemical transition state model with machine learning to predict the final outcomes of the reaction. Through quantum chemical calculations of the reaction mechanism and sensitivity analysis of important descriptors, we identify the transition state features of this reaction, which can aid future in-depth investigations. Our optimal machine learning model, the Support Vector Machine (SVM) [35,36], exhibits the highest predictive accuracy and demonstrates excellent precision and performance through subsequent experimental validation. This approach provides accurate guidance for scientists in designing and selecting optimal reaction conditions and holds promise for identifying other optimal reaction catalysts.

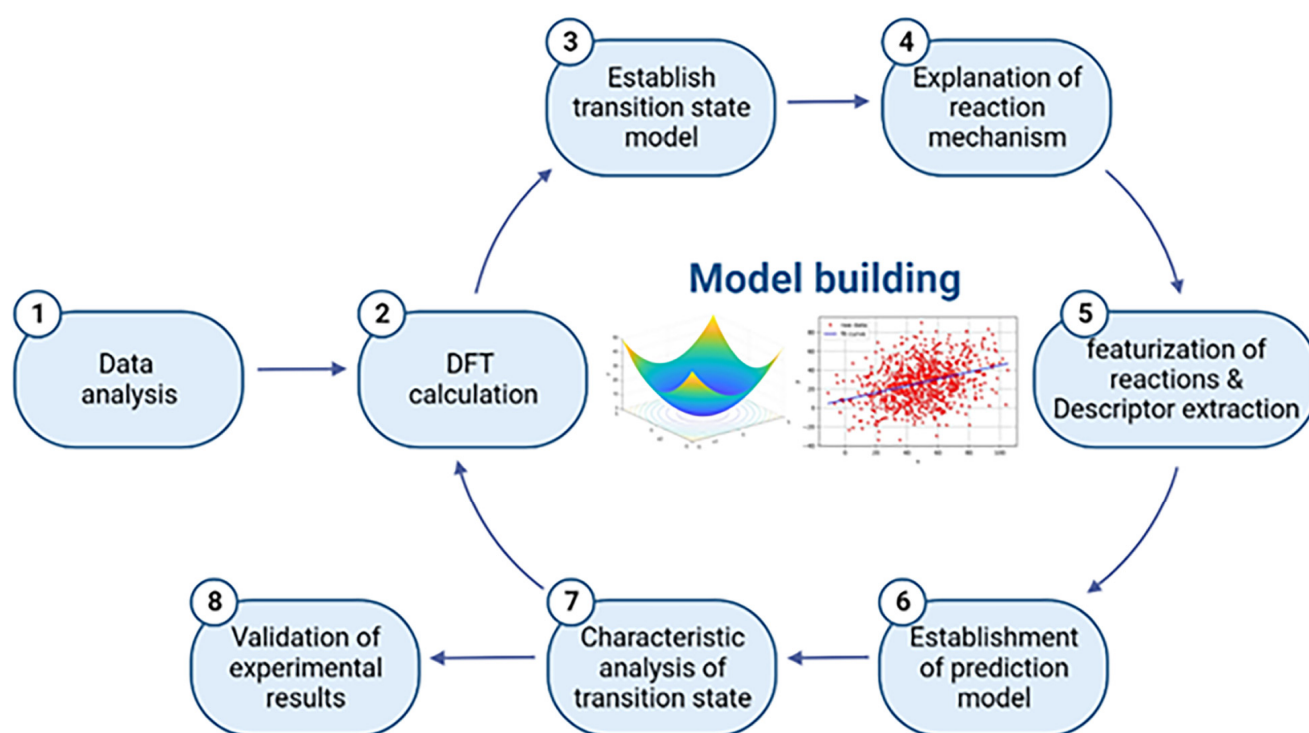


Figure 1. Flowchart of the machine learning methodology employed in this study.

2. Results and Discussion

A total of 110 experimental data on copper-catalyzed P–H insertion reactions were initially obtained from relevant literature. However, before extracting descriptors that accurately summarize catalyst performance, it is necessary to utilize density functional theory (DFT) to calculate and determine the transition state and reaction mechanism. Specifically, in the X–H insertion reaction of α -imino copper carbenes, it is important to understand why the reaction pathway is more inclined towards the 1,3-insertion pathway rather than the 1,1-insertion pathway. The literature extensively reports the DFT study

of the reaction process, where diazo compounds are converted to metal carbenes under a metal catalyst [37–41].

In this study, we utilized our newly synthesized catalyst, $\text{Cu}(\text{CH}_3\text{CN})_4\text{PF}_6$, as an example. As depicted in Figure 2a, the reaction barrier of this process aligns with the optimal reaction temperature of 50°C , further confirming that the intermediate state of the reaction is the α -imino copper carbene.

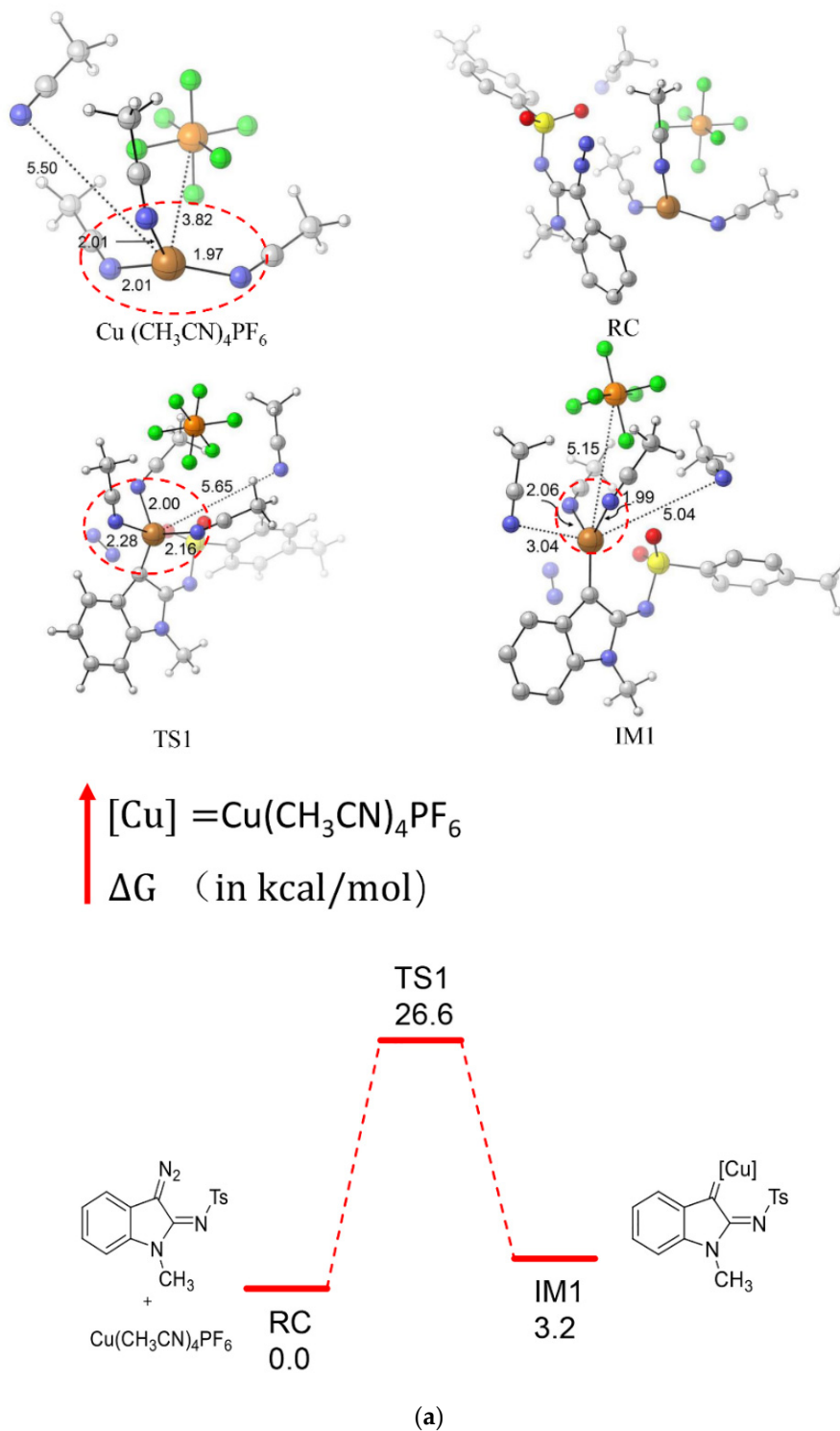


Figure 2. Cont.

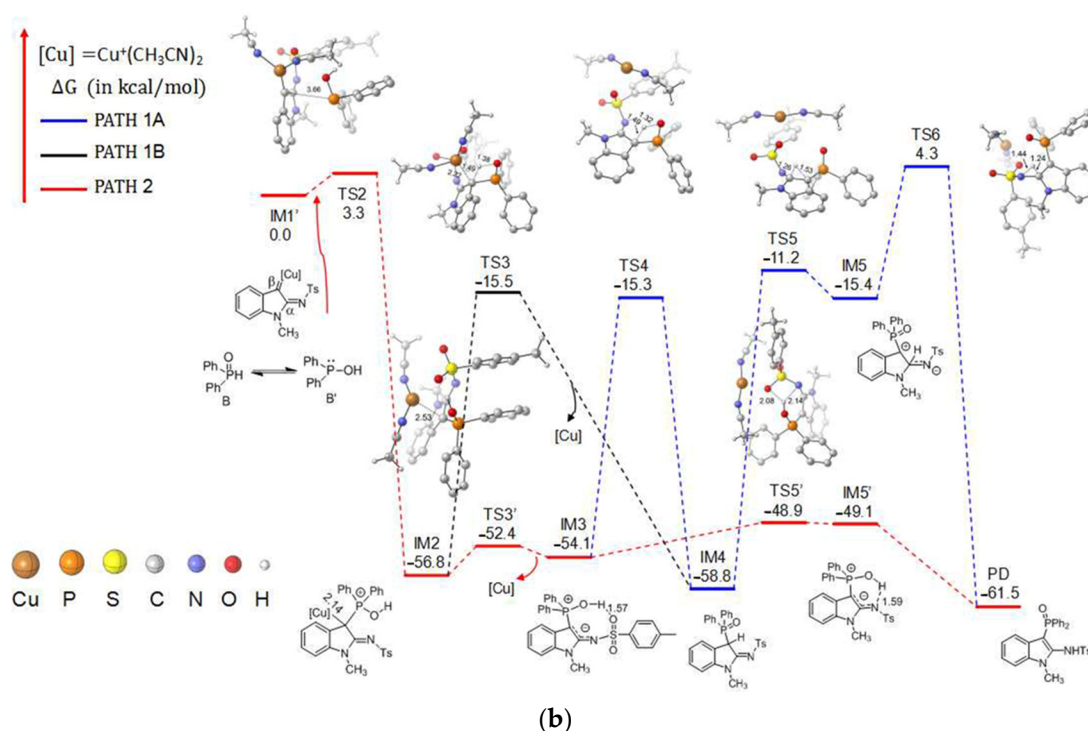


Figure 2. (a) (1) Gibbs free energy (in kcal/mol) profile obtained at the $\text{SMD}_{\text{CHCl}_3}/\text{M06}/6\text{-}311++\text{G}(\text{d},\text{p})\text{-LANL2DZ}/\text{B3LYP-D3}/6\text{-}31\text{G}(\text{d})\text{-LANL2DZ}$ level of theory for the formation of the α -imino copper carbene from 3-diazoindolin-2-imine catalyzed by $\text{Cu}(\text{CH}_3\text{CN})_4\text{PF}_6$. (2) Distance was given in Å. (3) Temperature of 323.15 K and a pressure of 1 atm. (b) (1) Gibbs free energy (in kcal/mol) profile obtained at the $\text{SMD}_{\text{CHCl}_3}/\text{M06}/6\text{-}311++\text{G}(\text{d},\text{p})\text{-LANL2DZ}/\text{B3LYP-D3}/6\text{-}31\text{G}(\text{d})\text{-LANL2DZ}$ level of theory for the formation of the 3-phosphonylindole product from α -imino copper carbene and *H*-phosphine oxides. The pathways shown in black and blue were 1,1-insertion reactions followed by hydrogen transfer processes, and the red pathway was a 1,3-insertion pathway. (2) Important intermediates and transition states considered in the computation of the energetic span for the catalytic cycle. (3) Distances were given in Å. (4) Temperature of 323.15 K and a pressure of 1 atm.

To gain a better understanding of the reaction mechanism, we performed Density Functional Theory (DFT) calculations on three hypothetical pathways, as depicted in Figure 2b. In PATH 1A (represented by the blue line in Figure 2b), the interaction between diphenylphosphinic acid and copper carbene intermediate IM1 leads to the formation of intermediate IM1'. Subsequently, IM1' proceeds to the copper-related intermediate IM2 via transition state TS2 with an activation energy barrier of only 3.3 kcal/mol. Once IM2 is formed, the copper catalyst dissociates through transition state TS3' with a ΔG of 4.4 kcal/mol, resulting in the production of the free ylide IM3. This indicates a high likelihood of the copper catalyst departure. In TS3', the $\beta\text{-C-Cu}$ distance changes from 2.14 Å in IM2 to 2.53 Å, while the $\beta\text{-C-Cu}$ distance in IM3 is 5.50 Å, indicating complete dissociation of the copper catalyst upon formation of IM3. Next, the proton migrates from diphenylphosphinic acid 2-2a' to the $\beta\text{-C}$ position of IM3 via transition state TS4, resulting in the formation of IM4 with an energy barrier of 38.8 kcal/mol. Subsequently, IM4 undergoes transformation into IM5 through intramolecular proton transfer, with an energy barrier of 47.6 kcal/mol, via TS5 from $\beta\text{-C}$ to $\alpha\text{-C}$. Finally, the proton is transferred from the $\alpha\text{-C}$ position of IM5 to the N atom of the Schiff base group, leading to the formation of the 3-phosphorylindole product (PD), with an energy barrier of 19.7 kcal/mol.

Compared to PATH 1A, PATH 1B (indicated by the black line in Figure 2b) differs only in the simultaneous occurrence of proton transfer between diphenylphosphinic acid 2-2a' and $\beta\text{-C}$ and the dissociation of the copper catalyst after the formation of IM2 via transition

state TS3. This results in the formation of IM4 with an energy barrier of 41.3 kcal/mol. PATH 1A and 1B explore the 1,1-insertion pathway of α -imino carbenes in P–H insertion reactions. However, the high activation energy barrier (47.6 kcal/mol) observed during the proton transfer process indicates that the 1,1-insertion pathway is unfavorable for the reaction. Additionally, this high activation energy barrier contradicts the reported reaction temperature of 50 °C in the literature. Thus, there must be another more reasonable reaction pathway for the P–H insertion reaction of α -imino carbenes.

On the other hand, PATH 2 (represented by the red line in Figure 2b) corresponds to the 1,3-insertion pathway for the copper-catalyzed P–H insertion reaction of α -imino carbenes. The process of forming the copper-related intermediate IM3 in PATH 2 is similar to that in PATH 1A. Through transition state TS2, diphenylphosphinic acid 2-2a' interacts with copper carbene intermediate IM1 to yield the copper-related intermediate IM2 ($\Delta G = 3.3$ kcal/mol). Once IM2 is formed, the copper catalyst dissociates through transition state TS3' with a $\Delta G = 4.4$ kcal/mol, producing the free ylide IM3. In IM3, there exists a strong hydrogen bond interaction (1.57 Å) between the hydroxyl group of the phosphinic acid and the oxygen atom of the sulfonyl group. Subsequently, in the presence of the Schiff base group, the hydroxyl group of the phosphinic acid forms a hydrogen bond (1.59 Å) with the nitrogen atom of the Schiff base group on the α -imino carbene through transition state TS5' (5.2 kcal/mol). In the case of the existing Schiff base, the O–H group of phosphinous acid tended to form a hydrogen bond with the N atom of the Schiff base group in IM5' via transition state TS5' (5.2 kcal/mol). Consequently, the proton of the phosphinous acid could be captured by the Schiff base through transition state TS6' (−1.1 kcal/mol) and finally generate 3-phosphonylindole (PD). This step played a key role in the 1,3-insertion pathway. This reaction pathway is deemed the most probable mechanism for the copper-catalyzed P–H insertion reaction of α -imino carbenes.

Although the DFT model for transition states is accurate, it remains a challenging task to determine the yield of catalytic reactions solely based on this model. Consequently, we have decided to integrate the transition state model with AI methods to facilitate the determination of the transition state characteristics and the prediction of the reaction yield. Previous reports combining quantum chemical transition state models with machine learning analysis to predict the yield of copper-catalyzed P–H insertion reactions are scarce. One important clarification needs to be made: the descriptors we obtained were based on calculations performed using α -amino copper carbenes as the transition state, as the various indices in this step are crucial in determining the feasibility of the reaction. After identifying the reactive transition state, we extracted 16 atomic and molecular descriptors from the transition state model using quantum chemical calculations. It is worth noting that in order to obtain more accurate calculations and more persuasive results, the experimental impact brought by solvent effects was taken into consideration during the establishment of the transition state model. However, in the subsequent machine learning modeling process, it was found that the solvent effects had little influence on the final prediction results. Moreover, considering the computational cost and time consumption in quantifying parameters, it was ultimately decided not to further consider solvent effects. These descriptors, along with the reaction yield, were utilized as the input and output datasets for our machine learning model. These descriptors, which may potentially impact the experimental catalytic yield, encompass the catalyst and reaction molecular mass (Mass), lipophilicity (Log P), water solubility (Log S), transition state energy (E-RB3LYP), dipole moment, polarizability, heat of formation (E (Thermal)), heat capacity, entropy, lowest occupied molecular orbital (LUMO), highest occupied molecular orbital (HOMO), length of Cu–C bond (Length (Cu–C)), length of P–C bond (Length (P–C)), Mulliken charge of Cu (Mulliken (Cu)), Mulliken charge of P (Mulliken (P)), and Mulliken charge of C (Mulliken (C)). Detailed calculation results for all data descriptors can be found in SI 1. To predict the performance and transition state characteristics of Cu catalysts, we employed five machine learning models: Partial Least Squares Regression (PLSR); Multiple Linear Regression (MLR); Stepwise Multiple Linear Regression (SMLR); Artificial Neural Networks (ANN);

and Support Vector Machine Regression (SVM). Each machine learning model provides prediction results based on its inherent algorithm, and comprehending the differences among them is crucial for selecting the most appropriate model for practical applications. In the field of computational chemistry and cheminformatics, the SVM is widely used for identifying new active compounds. Additionally, Support Vector Regression (SVR) has emerged as the preferred method for modeling non-linear structure–activity relationships and predicting compound potencies [42–46]. In our study, the SVM demonstrated the highest prediction accuracy among the five tested machine learning models, as well as the best performance in cross-validation (Figure 3). Therefore, we opted to utilize the SVM for further analysis.

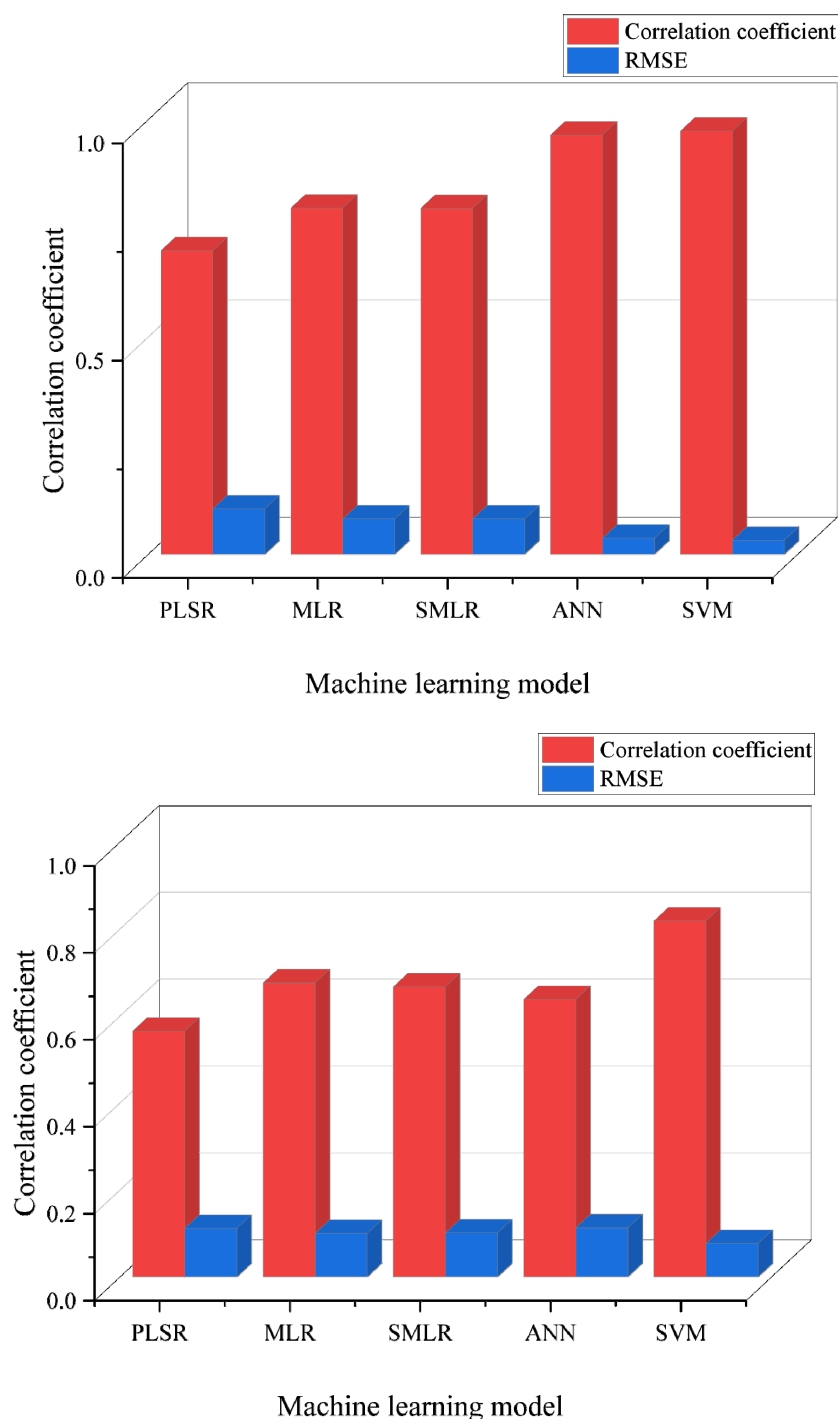


Figure 3. Regression coefficients and cross-validation results of various machine learning models.

After selecting the Support Vector Machine (SVM) as the optimal model, we further investigated the critical features that significantly affect the catalytic efficiency using the Principal Component Analysis (PCA) method, as illustrated in Figure 4. PCA is a multivariate statistical technique [47] that examines the correlations among multiple variables and explores how to reveal the internal structure of these variables by deriving a few principal components. These components preserve as much information as possible from the original variables while being uncorrelated. The identified descriptors, including length (Cu-C), Mass, Log P, and E (Thermal), align with expert chemists' intuitions regarding the relative importance of catalyzing P-H insertion reactions. The graph in Figure 4 clearly demonstrates that these descriptors have regression coefficients exceeding 1. Traditional analytical methods face challenges in quantifying these parameters accurately, but the combination of quantum chemistry calculations and machine learning provides a straightforward and direct approach.

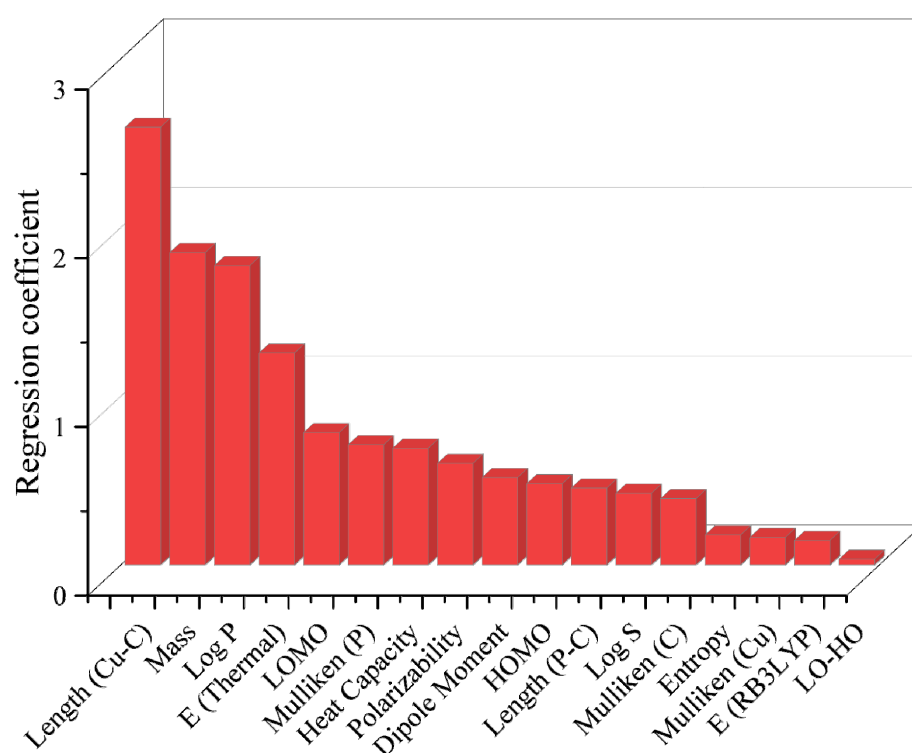


Figure 4. Importance plots of 16 descriptors.

To investigate the effect of the number of descriptors on predictive performance, we trained SVM models based on the top 16, 13, 10, and 7 descriptors shown in Figure 5. This demonstrated that reducing the number of descriptors from 16 to 13 has a negligible impact on predictive accuracy, and it further suggests that even when considering only the top 13 descriptors, the SVM model can provide satisfactory predictive results. Moreover, this indicates that the initially considered and selected descriptors are representative and accurately capture the factors influencing catalytic reaction outcomes.

The descriptors were obtained from quantitative calculations, which encompass a wide range of categories necessary for predicting the yield of the P-H insertion reaction. However, this also increases the complexity and computational time of the calculations. Additionally, due to the “black-box” nature of machine learning algorithms, we aim to establish a connection between this work and practical applications by providing interpretability rather than just predictive capabilities. Therefore, we conducted an importance analysis of these descriptors, gradually reducing their number and sequentially modeling them to observe their impact on prediction accuracy in order to gain a deeper understanding of these descriptors.

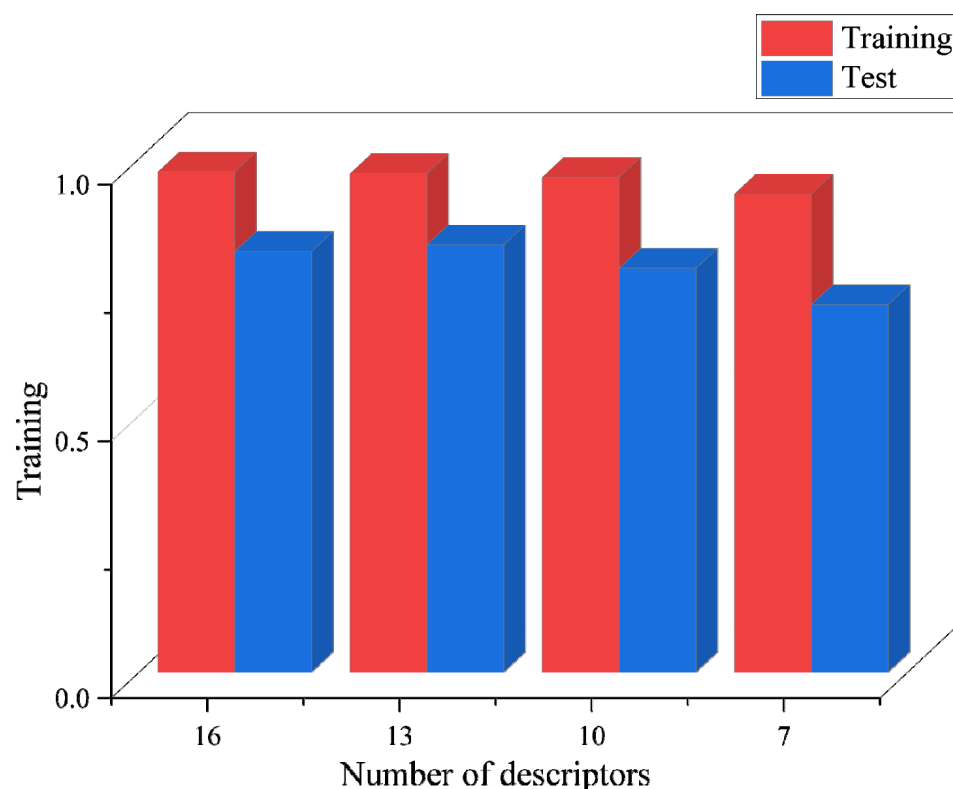


Figure 5. Model accuracy for different numbers of descriptors.

It can be observed that although the top four important descriptors contribute significantly to the overall importance, the prediction accuracy of the model gradually decreases as the number of descriptors is reduced. Even when only the top seven descriptors are used, the model can still achieve a cross-validation regression coefficient close to 0.8. However, this does not imply that descriptors other than the top-ranked ones are unimportant; their combination provides higher predictive accuracy. Our work offers a choice for situations where extensive calculations are not feasible, but an approximate estimation of reaction yield is needed. In such cases, only the top few important descriptors can be selected for modeling.

Further examination of Figure 6 reveals that length (Cu–C), Mass, Log P, E (Thermal), LOMO, and Mulliken (P) are the most influential factors affecting catalytic yield. To gain additional chemical insights, we obtained sensitivity plots for these descriptors from the SVM model, as shown in Figure 6. These plots display the catalytic yield as a function of descriptor variation. The results indicate that the reaction yield increases with the increase in Log P, E (Thermal), LOMO, and Mulliken (P) descriptors, while it decreases with the increase in length (Cu–C) of the molecule. The highest yield is observed at a molecular weight of approximately 500, which aligns with the expected trends and insights of chemists. The purpose of conducting sensitivity analyses is to establish a connection with real-world experiments, enhancing interpretability and providing guidance for the further exploration of reactions. For instance, in the P–H insertion reaction, an increase in Cu–C distance implies a decrease in the likelihood of the reaction occurrence. The molecular weight of the transition state should fall within an appropriate range to facilitate the reaction, as deviations towards larger or smaller values can lead to a decrease in yield. Given that most reactions occur in organic solvents, a stronger lipophilicity is associated with higher yields. Hence, it is important to consider the lipophilicity of reactants in subsequent experimental processes. However, despite the intuitive nature of variables such as molecular weight and lipophilicity, other descriptors, though informative in understanding their significance and impact on yield variations, are not easily controlled, posing limitations in experimental design.

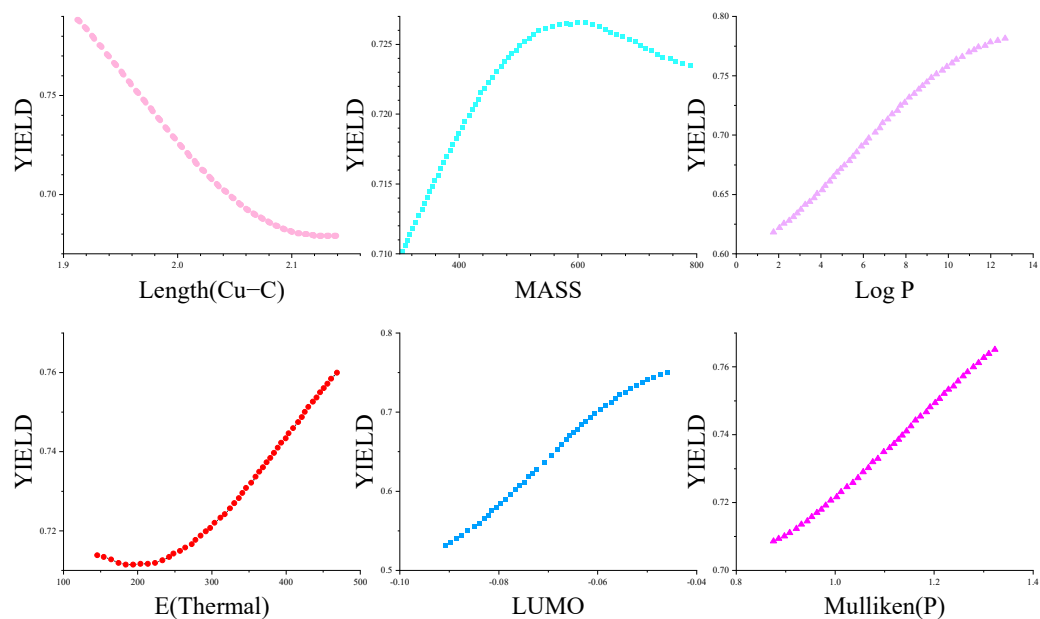


Figure 6. Sensitivity graph of each descriptor.

To demonstrate the reliability of the constructed model in a more intuitive manner, we conducted experiments on 26 synthetic design schemes, and the final experimental results and model prediction curves are shown in SI 2 and Figure 7. Among these schemes, the pink squares and blue circles represent samples with yields $>80\%$ and $\leq 80\%$, respectively, collected from other literature sources. The red inverted triangles and blue triangles represent 26 samples with yields $>80\%$ and $\leq 80\%$, respectively, obtained in this experiment. Although slight discrepancies exist between the experimental and calculated values, mainly due to differences in experimental conditions and idealization of the simulation model, Figure 7 demonstrates that the accuracy and reliability of our established prediction model can be further verified through experiments.

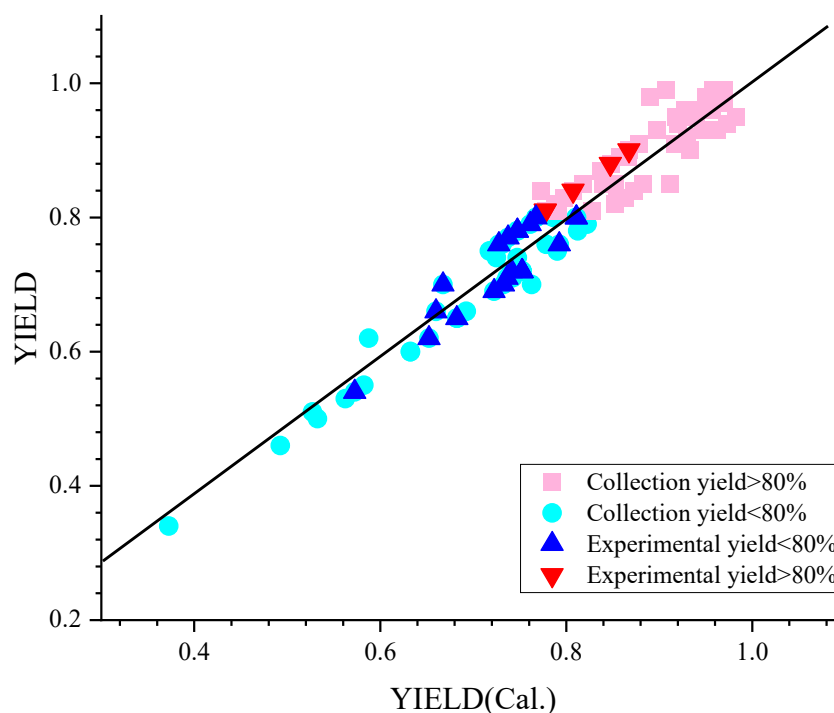


Figure 7. Graph of experimental results and predicted curve.

3. Experimental Section

3.1. Data Source

All the data used for modeling in this paper were sourced from five publicly available literature publications that discuss copper-catalyzed P–H insertion reactions [48–52].

3.2. Quantum Chemistry Calculations and Descriptor Acquisition

All theoretical calculations in this chapter were conducted using the Gaussian 16 software package [53] based on density functional theory (DFT). The B3LYP-D3 density functional [54–58] was employed for geometry optimization of all reaction stationary points. The metal copper atoms were modeled using the LANL2DZ [59,60] pseudo-potential basis set, while the 6-31G(d) basis set [61] was used for all other atoms. Frequency calculations were performed to determine whether the stationary points corresponded to minimum values or first-order saddle points. In addition, intrinsic reaction coordinate (IRC) calculations were carried out at the same theoretical level [62–64] to confirm the connectivity between the relevant reactants and products, thereby verifying the accuracy of the transition state.

For all single-point energy calculations, the optimized structures obtained from the geometry optimization were used as a basis at the B3LYP-D3/6-31G(d)-LANL2DZ level. Furthermore, the 6-311++G(d,p) basis set [65,66] was utilized for all atoms except copper, and the M06 algorithm [67] was employed for these calculations. To account for solvent effects, the Truhlar and Cramer-developed SMD solvent model [68] was employed. The solvent correction was performed using the SMDCHCl₃/M06/6-311++G(d,p)-LANL2DZ theoretical level. The Gibbs free energy was calculated at a temperature of 323.15 K and a pressure of 1 atm, based on the actual reaction temperature of 50 °C.

3.3. Machine Learning Models

The model construction and testing in this study were carried out using analytical software (ExMiner 1.8.7.8) developed by our laboratory [69]. The ML predictions rely on the selection of algorithms, and even experienced data scientists cannot determine the best-performing algorithm without experimenting with different ones. Hence, in this study, five ML models were established utilizing preprocessed datasets.

3.3.1. Partial Least Squares Regression (PLSR)

The principle behind PLSR [70] is to find a linear regression model by projecting the predictor variables and the observed variables onto a new space through.

3.3.2. Multiple Linear Regression (MLR)

The basic principle of MLR [71] is similar to that of simple linear regression, with the difference being the involvement of two or more independent variables.

3.3.3. Stepwise Multiple Linear Regression (SMLR)

SMLR [72] analysis involves the gradual introduction of variables, with each new variable being tested against the previously selected variables to ensure that each variable in the resulting subset is significant. The process is repeated until no further variables can be added.

3.3.4. Artificial Neural Networks (ANN)

ANN [73] is a computational model composed of a large number of interconnected nodes (or neurons). Each node represents a specific output function, known as an activation function. The connection between any two nodes represents a weighted value for the signal passing through that connection, known as a weight. Artificial Neural Networks simulate human memory through this mechanism. The output of the network depends on its structure, connection pattern, weights, and activation functions.

3.3.5. Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a powerful machine learning algorithm that is widely used in classification and regression tasks. The core concept of the SVM is to find a hyperplane that maximizes the margin between two classes of data, allowing for effective classification. The choice of kernel function is crucial in the SVM, as it determines how the data are transformed and classified [74].

In regression tasks, Support Vector Machine Regression (SVR) is a significant application of the SVM. SVR aims to find a regression plane that minimizes the distance between all data points and the plane. This approach allows for the accurate prediction and modeling of continuous variables.

To apply the SVM to regression, an alternative loss function is introduced. The results obtained via SVR have shown promising performance. The key idea behind SVR is to map the input data X into a higher-dimensional feature space F using a non-linear mapping function Φ . Regression is then performed in this transformed space, enabling the modeling of complex relationships between variables.

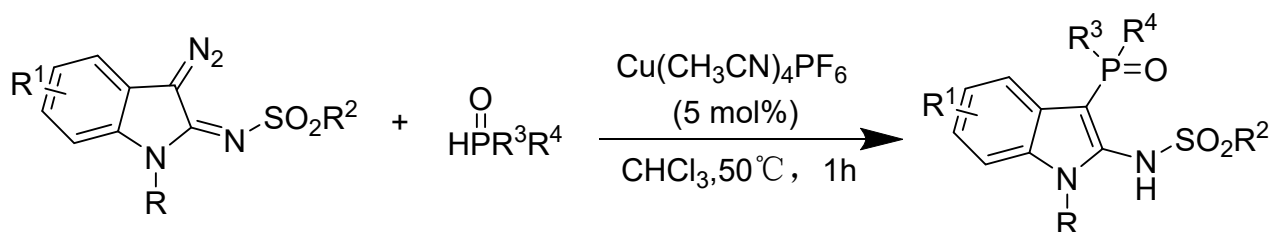
In practical applications, non-linear models are often necessary for better data fitting. Similar to the non-linear support vector classification approach, non-linear mapping can be employed to transform the data into a higher-dimensional feature space. In this transformed space, linear regression can be applied to accurately model the data.

The complete SVM algorithm can be described in terms of dot products between data points. The dot product measures the similarity between two data points and is used to determine the position of the hyperplane that separates the different classes. By maximizing the margin between classes, the SVM ensures robust classification and regression results.

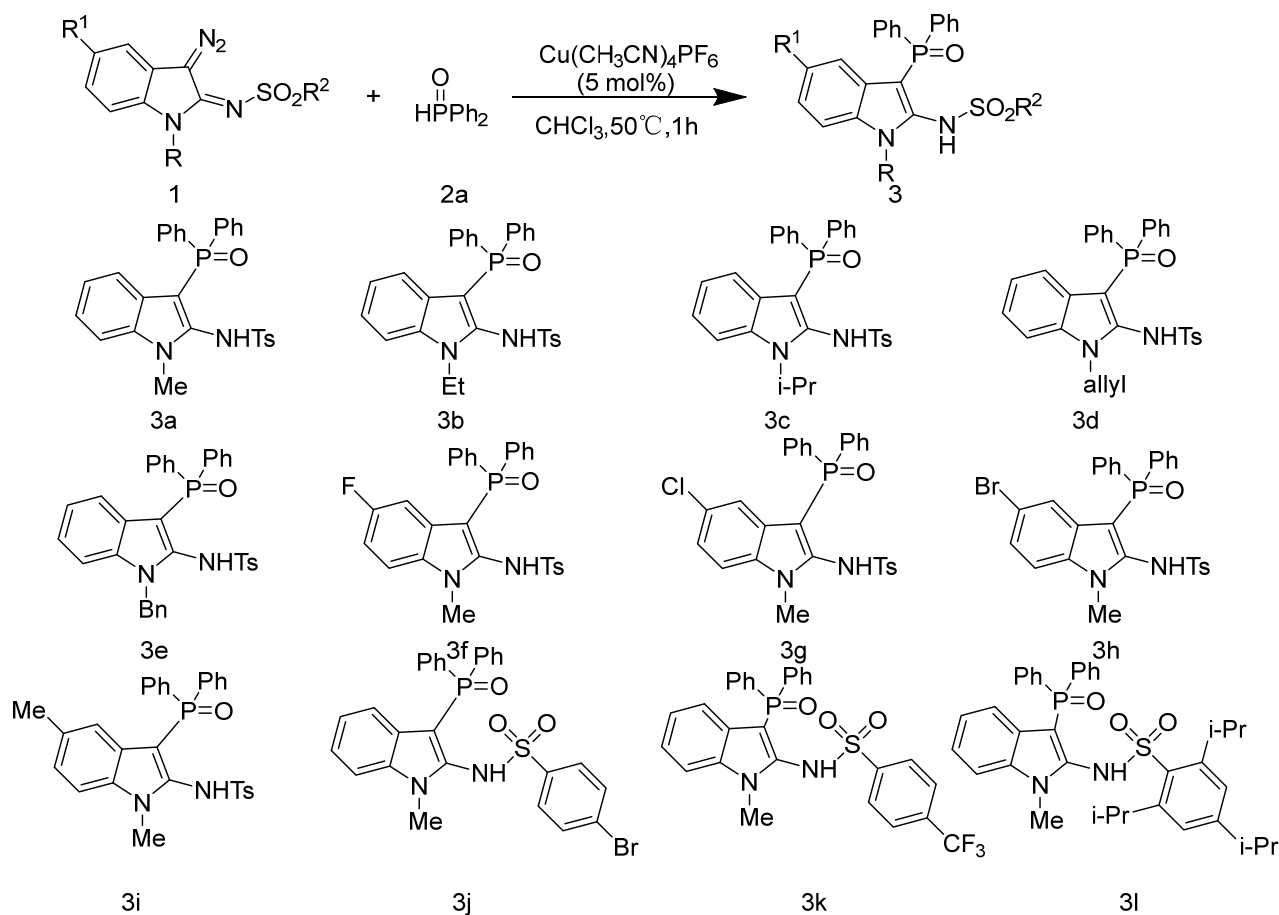
In conclusion, the Support Vector Machine is a versatile algorithm that can be used for both classification and regression tasks. SVM regression (or SVR) allows for the accurate prediction of continuous variables by finding a regression plane that minimizes the distance to the data points. Non-linear models can be achieved through the use of kernel functions and by mapping the data into a higher-dimensional feature space. The dot product between data points plays a crucial role in determining the position of the hyperplane and achieving optimal classification and regression results.

3.4. Synthesis

Based on previous research conducted in our laboratory, a reaction scheme for copper-catalyzed P–H insertion was devised, as depicted in Schemes 1–4. The design and discussion encompassed the substitution of both the 1st and 5th positions of the indole substrate, along with various sulfonyl groups. Moreover, the different types of H-type phosphine oxides and the subsequent modifications in experimental outcomes resulting from the addition of chiral reagents were taken into consideration. To begin the experiment, a mixture of 3-azidoindole-2-imine (0.2 mmol) and 0.4 mL of CHCl_3 was gradually added to a mixed solution containing H-type phosphine oxide (0.2 mol), catalyst $\text{Cu}(\text{CH}_3\text{CN})_4\text{PF}_6$ (0.01 mmol), and 0.4 mL of CHCl_3 . The reaction mixture was then stirred at 50°C under an argon atmosphere for one hour. Upon completion, the solvent was evaporated under vacuum, and the resulting residue was purified using silica gel column chromatography (petroleum ether/ethyl acetate 3:1) to obtain the phosphine hydride compound.



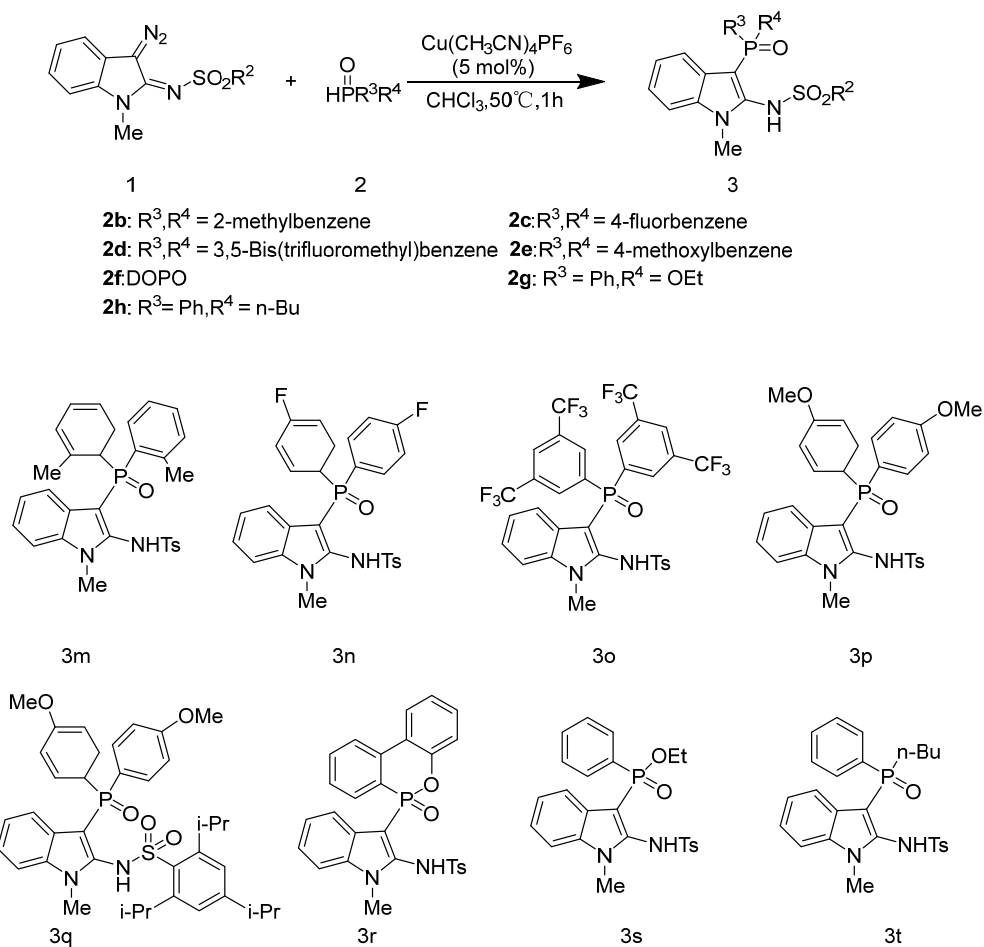
Scheme 1. Overall design scheme of the P–H insertion reaction.



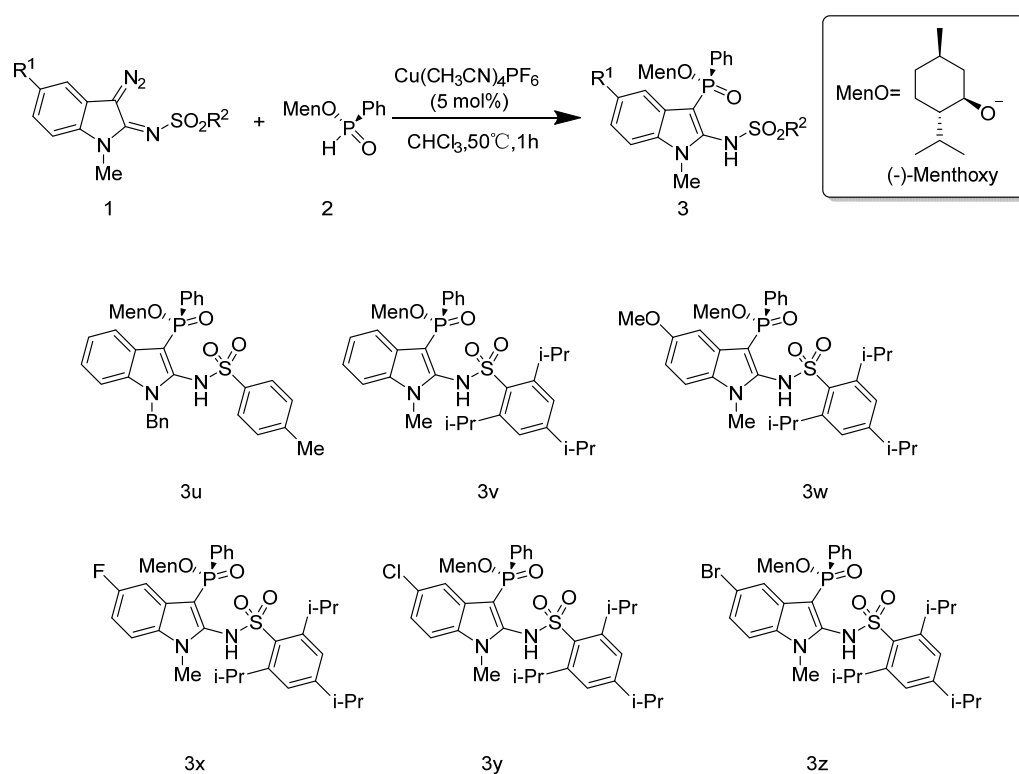
Scheme 2. Partial design scheme of the P-H insertion reaction.

3.5. Structural Validation of Synthesized Product

The ^1H NMR and ^{13}C NMR spectra were acquired using a Bruker 600 MHz spectrometer in CDCl_3 . TMS served as the internal standard for ^1H NMR ($\delta = 0$), while CDCl_3 was employed as the internal standard for ^{13}C NMR ($\delta = 77.0$). Additionally, the ^{31}P NMR and ^{19}F NMR spectra were recorded on the same instrument. Chemical shifts were reported in parts per million (ppm), and the multiplicity was indicated as s (singlet), d (doublet), t (triplet), q (quartet), m (multiplet), or br (broad). High-resolution mass spectrometry (HRMS) using electrospray ionization (ESI) was performed on a Thermo Fisher Scientific LTQ FT Ultra. The starting materials were purchased from Aldrich, Macklin, and Energy Chemicals and were used without further purification. Solvents were dried and purified following the procedures. Column chromatography was conducted using silica gel (200–300 mesh ASTM). The substrates were prepared according to published procedures [75].



Scheme 3. Partial design scheme of the P-H insertion reaction.



Scheme 4. Partial design scheme of the P-H insertion reaction.

4. Conclusions

This work presents a novel approach that combines quantum chemical transition state modeling with machine learning to establish a highly accurate model for predicting transition state features and yield in P–H insertion reactions. This study proves the potential of integrating quantum mechanical calculations and machine learning techniques to predict the outcome of catalytic reactions, which could significantly reduce the costs of human labor and experimentation. Furthermore, by developing appropriate descriptors and fine-tuning hyperparameters, this method can be extended to other organic and inorganic material fields, thereby facilitating the improvement and discovery of new materials. Furthermore, it is important to recognize that there is still room for further improvement in our work—for example, exploring the inclusion of solvent-related descriptors to enhance the general applicability and stability of the prediction model. Additionally, conducting sensitivity analyses on different descriptors would provide theoretical guidance for subsequent experimental designs, rather than relying solely on designing predictions before conducting experiments. This approach would enhance the practicality and functionality of the model, and it is an area we will explore in our future investigations.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/molecules28165995/s1>

Author Contributions: Methodology, X.F.; Software, X.F.; Validation, L.Z.; Formal analysis, Y.M.; Investigation, J.J.; Resources, L.W. and L.J.; Data curation, Y.M., X.Z. and J.A.H.K.; Writing—original draft, Y.M.; Writing—review & editing, X.L.; Supervision, L.W. and X.L.; Project administration, X.L.; Funding acquisition, L.J. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All relevant data generated during our research process have been uploaded to the Supplementary Files.

Conflicts of Interest: The authors declare no conflict of interest.

Sample Availability: Not applicable.

References

1. Monai, M.; Weckhuysen, B.M. Crowded catalyst, better catalyst. *Natl. Sci. Rev.* **2021**, *8*, nwab141. [CrossRef]
2. Lim, X. The new breed of cutting-edge catalysts. *Nature* **2016**, *537*, 156–158. [CrossRef]
3. Dai, L. Metal-Free Carbon Electrocatalysts: Recent Advances and Challenges Ahead. *Adv. Mater.* **2019**, *31*, e1900973. [CrossRef]
4. Zhu, S.F.; Zhou, Q.L. Transition-Metal-Catalyzed Enantioselective Heteroatom-Hydrogen Bond Insertion Reactions. *Acc. Chem. Res.* **2012**, *45*, 1365–1377. [CrossRef]
5. Bergstrom, B.D.; Nickerson, L.A.; Shaw, J.T.; Souza, L.W. Transition Metal Catalyzed Insertion Reactions with Donor/Donor Carbenes. *Angew. Chem.* **2020**, *133*, 6940–6954. [CrossRef]
6. Batista, V.F.; GA Pinto, D.C.; Silva, A.M. Iron: A Worthy Contender in Metal Carbene Chemistry. *ACS Catal.* **2020**, *10*, 10096–10116. [CrossRef]
7. Candeias, N.R.; Paterna, R.; Gois, P. Homologation Reaction of Ketones with Diazo Compounds. *Chem. Rev.* **2016**, *116*, 2937–2981. [CrossRef]
8. Wang, Y.F.; Wang, C.J.; Feng, Q.Z.; Zhai, J.J.; Qi, S.S.; Zhong, A.G.; Chu, M.M.; Xu, D.Q. Copper-catalyzed asymmetric 1,6-conjugate addition of in situ generated para-quinone methides with β -ketoesters. *Chem. Commun.* **2022**, *58*, 6653–6656. [CrossRef]
9. Doyle, M.P.; McKervey, M.A.; Ye, T. *Modern Catalytic Methods for Organic Synthesis with Diazo Compounds (From Cyclopropanes to Ylides)*; Wiley-Interscience: New York, NY, USA, 1998; 652p, ISBN 0-47113556-9.
10. Hosseinian, A.; Farshbaf, S.; Fekri, L.Z.; Nikpassand, M.; Vessally, E. Cross-Dehydrogenative Coupling Reactions Between P(O)-H and X-H (X = S, N, O, P) Bonds. *Top. Curr. Chem.* **2018**, *376*, 23. [CrossRef]
11. Zhang, X.; Zhang, Y.; Liang, C.; Jiang, J. Copper-catalyzed P-H insertion reactions of sulfoxonium ylides. *Org. Biomol. Chem.* **2021**, *19*, 5767–5771. [CrossRef]

12. Wu, Y.; Chen, K.; Ge, X.; Ma, P.; Xu, Z.; Lu, H.; Li, G. Redox-Neutral P(O)-N Coupling between P(O)-H Compounds and Azides via Dual Copper and Photoredox Catalysis. *Org. Lett.* **2020**, *22*, 6143–6149. [[CrossRef](#)]
13. Zhou, Y.; Yin, S.; Gao, Y.; Zhao, Y.; Goto, M.; Han, L.B. Selective P-P and P-O-P bond formations through copper-catalyzed aerobic oxidative dehydrogenative couplings of H-phosphonates. *Angew. Chem. Int. Ed. Engl.* **2010**, *49*, 6852–6855. [[CrossRef](#)] [[PubMed](#)]
14. Ess, D.; Gagliardi, L.; Hammes-Schiffer, S. Introduction: Computational Design of Catalysts from Molecules to Materials. *Chem. Rev.* **2019**, *119*, 6507–6508. [[CrossRef](#)] [[PubMed](#)]
15. Wang, Y.; Wang, J.; Su, J.; Huang, F.; Jiao, L.; Liang, Y.; Yang, D.; Zhang, S.; Wender, P.A.; Yu, Z.-X. A Computationally Designed Rh(I)-Catalyzed Two-Component [5+2+1] Cycloaddition of Ene-vinylcyclopropanes and CO for the Synthesis of Cyclooctenones. *J. Am. Chem. Soc.* **2007**, *129*, 10060–10061. [[CrossRef](#)] [[PubMed](#)]
16. Donoghue, P.J.; Helquist, P.; Norrby, P.-O.; Wiest, O. Prediction of Enantioselectivity in Rhodium Catalyzed Hydrogenations. *J. Am. Chem. Soc.* **2009**, *131*, 410–411. [[CrossRef](#)]
17. Rowley, C.N.; Woo, T.K. Computational design of ruthenium hydride olefin-hydrogenation catalysts containing hemilabile ligands. *Can. J. Chem.* **2009**, *87*, 1030–1038. [[CrossRef](#)]
18. Baik, M.-H.; Mazumder, S.; Ricci, P.; Sawyer, J.R.; Song, Y.-G.; Wang, H.; Evans, P.A. Computationally Designed and Experimentally Confirmed Diastereoselective Rhodium-Catalyzed Pauson–Khand Reaction at Room Temperature. *J. Am. Chem. Soc.* **2011**, *133*, 7621–7623. [[CrossRef](#)]
19. Fernandez, L.E.; Horvath, S.; Hammes-Schiffer, S. Theoretical Design of Molecular Electrocatalysts with Flexible Pendant Amines for Hydrogen Production and Oxidation. *J. Phys. Chem. Lett.* **2013**, *4*, 542–546. [[CrossRef](#)]
20. Nielsen, M.C.; Bonney, K.J.; Schoenebeck, F. Computational Ligand Design for the Reductive Elimination of ArCF₃ from a Small Bite Angle PdII Complex: Remarkable Effect of a Perfluoroalkyl Phosphine. *Angew. Chem. Int. Ed.* **2014**, *53*, 5903–5906. [[CrossRef](#)]
21. Bernales, V.; League, A.B.; Li, Z.; Schweitzer, N.M.; Peters, A.W.; Carlson, R.K.; Hupp, J.T.; Cramer, C.J.; Farha, O.K.; Gagliardi, L. Computationally Guided Discovery of a Catalytic Cobalt-Decorated Metal–Organic Framework for Ethylene Dimerization. *J. Phys. Chem. C* **2016**, *120*, 23576–23583. [[CrossRef](#)]
22. Boddada, A.; Hossain, M.M.; Mirzaei, M.S.; Lindeman, S.V.; Mirzaei, S.; Rathore, R. Angular ladder-type meta-phenylenes: Synthesis and electronic structural analysis. *Org. Chem. Front.* **2020**, *7*, 3215–3222. [[CrossRef](#)]
23. Cao, X.; Rong, C.; Zhong, A.; Lu, T.; Liu, S. Molecular acidity: An accurate description with information-theoretic approach in density functional reactivity theory. *J. Comput. Chem.* **2018**, *39*, 117–129. [[CrossRef](#)] [[PubMed](#)]
24. Schumann, J.; Medford, A.J.; Yoo, J.S.; Zhao, Z.J.; Norskov, J.K. Selectivity of Synthesis Gas Conversion to C₂+ Oxygenates on fcc(111) Transition-Metal Surfaces. *ACS Catal.* **2018**, *8*, 3447–3453. [[CrossRef](#)]
25. Ulissi, Z.W.; Medford, A.J.; Bligaard, T.; Norskov, J.K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **2017**, *8*, 14621. [[CrossRef](#)]
26. Ahneman, D.T.; Estrada, J.G.; Lin, S.; Dreher, S.D.; Doyle, A.G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360*, eaar5169. [[CrossRef](#)] [[PubMed](#)]
27. O'Connor, N.; Jonayat, A.; Janik, M.J.; Senftle, T.P. Interaction trends between single metal atoms and oxide supports identified with density functional theory and statistical learning. *Nat. Catal.* **2018**, *1*, 531–539. [[CrossRef](#)]
28. Sun, M.; Wu, T.; Xue, Y.; Dougherty, A.W.; Yan, C.H. Mapping of atomic catalyst on graphdiyne. *Nano Energy* **2019**, *62*, 754–763. [[CrossRef](#)]
29. Wang, X.; Ye, S.; Hu, W.; Sharman, E.; Liu, R.; Liu, Y.; Luo, Y.; Jiang, J. Electric Dipole Descriptor for Machine Learning Prediction of Catalyst Surface–Molecular Adsorbate Interactions. *J. Am. Chem. Soc.* **2020**, *142*, 7737–7743. [[CrossRef](#)]
30. Tran, K.; Ulissi, Z.W. Active learning across intermetallics to guide discovery of electrocatalysts for CO₂ reduction and H₂ evolution. *Nat. Catal.* **2018**, *1*, 696–703. [[CrossRef](#)]
31. Huang, Y.; Chen, Y.; Cheng, T.; Wang, L.-W.; Goddard, W.A., III. Identification of the Selective Sites for Electrochemical Reduction of CO to C₂+ Products on Copper Nanoparticles by Combining Reactive Force Fields, Density Functional Theory, and Machine Learning. *ACS Energy Lett.* **2018**, *3*, 2983–2988. [[CrossRef](#)]
32. Sun, M.; Dougherty, A.W.; Huang, B.; Li, Y.; Yan, C.-H. Accelerating Atomic Catalyst Discovery by Theoretical Calculations–Machine Learning Strategy. *Adv. Energy Mater.* **2020**, *10*, 1903949. [[CrossRef](#)]
33. Artrith, N.; Lin, Z.; Chen, J.G. Predicting the Activity and Selectivity of Bimetallic Metal Catalysts for Ethanol Reforming using Machine Learning. *ACS Catal.* **2020**, *10*, 9438–9444. [[CrossRef](#)]
34. Meyer, B.; Sawatlon, B.; Heinen, S.; Lilienfeld, O.; Corminboeuf, C. Machine learning meets volcano plots: Computational discovery of cross-coupling catalysts. *Chem. Sci.* **2018**, *9*, 7069–7077. [[CrossRef](#)] [[PubMed](#)]
35. Xu, L.; Wencong, L.; Chunrong, P.; Qiang, S.; Jin, G. Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO₃ perovskites. *Comput. Mater. Sci.* **2009**, *46*, 860–868. [[CrossRef](#)]
36. Chen, N.; Lu, W.; Yang, J.; Li, G. *Support Vector Machine in Chemistry*; World Scientific: Singapore, 2004.
37. Lvarez, M.; Besora, M.; Molina, F.; Maseras, F.; Pérez, P. Two Copper–Carbenes from One Diazo Compound. *J. Am. Chem. Soc.* **2021**, *143*, 4837–4843. [[CrossRef](#)] [[PubMed](#)]
38. Balhara, R.; Chatterjee, R.; Jindal, G. A computational approach to understand the role of metals and axial ligands in artificial heme enzyme catalyzed C–H insertion. *Phys. Chem. Chem. Phys.* **2021**, *23*, 9500–9511. [[CrossRef](#)]
39. Lübcke, M.; Szabó, K. Diazocarbonyl Compounds in Organofluorine Chemistry. *Synlett* **2020**, *32*, 1060–1071.

40. Wu, Y.; Cao, S.; Douair, I.; Maron, L.; Bi, X. Computational Insights into Different Mechanisms for Ag-, Cu-, and Pd-Catalyzed Cyclopropanation of Alkenes and Sulfonyl Hydrazones. *Chem. Eur. J.* **2021**, *27*, 5999–6006. [CrossRef]
41. Zhang, Y.; Yang, Y.; Zhao, J.; Xue, Y. Mechanism and Diastereoselectivity of [3+3] Cycloaddition between Enol Diazoacetate and Azomethine Imine Catalyzed by Dirhodium Tetracarboxylate: A Theoretical Study. *Eur. J. Org. Chem.* **2018**, *2018*, 3086–3094. [CrossRef]
42. Varnek, A.; Baskin, I. Machine learning methods for property prediction in chemoinformatics: Quo Vadis? *J. Chem. Inf. Model.* **2012**, *52*, 1413–1437. [CrossRef]
43. Vogt, M.; Bajorath, J. Chemoinformatics: A view of the field and current trends in method development. *Bioorg. Med. Chem.* **2012**, *20*, 5317–5323. [CrossRef]
44. Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5–14. [CrossRef]
45. Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216. [CrossRef]
46. Heikamp, K.; Bajorath, J. Support vector machines for drug discovery. *Expert Opin. Drug Discov.* **2014**, *9*, 93–104. [CrossRef] [PubMed]
47. Shlens, J. A Tutorial on Principal Component Analysis. *Int. J. Remote Sens.* **2014**, *51*. [CrossRef]
48. Shen, L.B. Copper(II) Acetate-Catalyzed Synthesis of Phosphorylated Pyridines via Denitrogenative C-P Coupling between Pyridotriazoles and P(O)H Compounds. *Adv. Synth. Catal.* **2018**, *360*, 4252–4258. [CrossRef]
49. Qian, C.; Xinxing, Y.; Chunxiao, W.; Jiekun, Z.; Yulin, H.; Xingguo, L.; Kun, Z. Copper-Catalyzed Addition of H-P(O) Bonds to Arynes. *J. Org. Chem.* **2016**, *81*, 9476–9482.
50. Lu, L.Q.; Wang, B.C.; Wang, Y.N.; Zhang, M.M.; Xiao, W.J. Copper-catalyzed decarboxylative cyclization via tandem C-P and C-N bond formation: Access to 2-phosphorylmethyl indoles. *Chem. Commun.* **2018**, *54*, 3154–3157.
51. Liu, X.-Y.; Zou, Y.-X.; Ni, H.-L.; Zhang, J.; Dong, H.-B.; Chen, L. Copper-catalyzed tandem phosphorylative allenylation/cyclization of 1-(o-aminophenyl)prop-2-ynols with the P(O)-H species: Access to C2-phosphorylmethylindoles. *Org. Chem. Front.* **2020**, *7*, 980–986. [CrossRef]
52. Shen, R.; Yang, J.; Luo, B.; Zhang, L.; Han, L.B. Copper-Catalyzed Allenylation-Isomerization Sequence of Penta-1,4-diyne-3-yl Acetates with P(O)H Compounds: Facile Synthesis of 1-Phosphonyl 2,4-Diynes. *Adv. Synth. Catal.* **2016**, *358*, 3897–3906. [CrossRef]
53. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.A. Gaussian 09, Revision B.01. 2010. Available online: <https://www.scienceopen.com/document?vid=45b5a7ba-f6ee-40ce-b346-7407f99a540d> (accessed on 30 June 2023).
54. Aleku, G.A.; Saaret, A.; Bradshaw-Allen, R.T.; Derrington, S.R.; Titchiner, G.R.; Gostimskaya, I.; Gahlloth, D.; Parker, D.A.; Hay, S.; Leys, D. Enzymatic C-H activation of aromatic compounds through CO₂ fixation fixation. *Nat. Chem. Biol.* **2020**, *16*, 1255–1260. [CrossRef] [PubMed]
55. Raghavachari, K. Perspective on “Density functional thermochemistry. III. The role of exact exchange”. *Theor. Chem. Acc.* **2000**, *103*, 361–363. [CrossRef]
56. Deng, Y.; Yu, D.; Cao, X.; Liu, L.; Rong, C.; Lu, T.; Liu, S. Structure, aromaticity and reactivity of corannulene and its analogues: A conceptual density functional theory and density functional reactivity theory study. *Mol. Phys.* **2017**, *116*, 956–968. [CrossRef]
57. Qi, D.; Zhang, L.; Wan, L.; Zhang, Y.; Bian, Y.; Jiang, J. Conformational effects, molecular orbitals, and reaction activities of bis(phthalocyaninato) lanthanum double-deckers: Density functional theory calculations. *Phys. Chem. Chem. Phys.* **2011**, *13*, 13277–13286. [CrossRef]
58. Stephens, P.J.; Devlin, F.J.; Chabalowski, C.F.; Frisch, M.J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 247–257. [CrossRef]
59. Hay, P.J.; Wadt, W.R. Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.* **1985**, *82*, 299–310. [CrossRef]
60. Roy, L.E.; Hay, P.J.; Martin, R.L. Revised Basis Sets for the LANL Effective Core Potentials. *J. Chem. Theory Comput.* **2008**, *4*, 1029. [CrossRef]
61. Hariharan, P.C.P.; Pople, J.A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta* **1973**, *28*, 213–222. [CrossRef]
62. Fukui, K. A Formulation of Reaction Coordinate. *J. Phys. Chem.* **1970**, *74*, 4161–4163. [CrossRef]
63. Fukui, K. The Path of Chemical Reactions—The IRC Approach. *Acc. Chem. Res.* **1981**, *14*, 363–368. [CrossRef]
64. Uddin, K.M.; Poirier, R.A. Computational Study of the Deamination of 8-Oxoguanine. *J. Phys. Chem. B* **2011**, *115*, 9151–9159. [CrossRef]
65. Krishnan, R.; Binkley, J.S.; Seeger, R.; Pople, J.A. Self consistent molecular orbital methods. A basis set for correlated wave functions. *J. Chem. Phys.* **1980**, *72*, 650–654.
66. Mclean, A.D.; Chandler, G.S. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18. *J. Chem. Phys.* **1980**, *72*, 5639–5648. [CrossRef]

67. Zhao, Y.; Truhlar, D.G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *119*, 525.
68. Marenich, A.V.; Cramer, C.J.; Truhlar, D.G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396. [[CrossRef](#)]
69. Xu, L.; Wencong, L.; Shengli, J.; Yawei, L.; Nianyi, C. Support Vector Regression Applied to Materials Optimization of SiAlON Ceramics. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 8–14. [[CrossRef](#)]
70. Lorber, A.; Wangen, L.E.; Kowalski, B.R. A Theoretical Foundation for the PLS Algorithm. *J. Chemom.* **1987**, *1*, 19–31. [[CrossRef](#)]
71. Bin, L. Multiple linear regression analysis and its application. *China Sci. Technol. Inf.* **2010**, *3*, 1–25.
72. Xia, Z.; Yu, B.; Yuan, X. ON ¹³C NMR SPECTROSCOPY: Approach to Chemical Shift Sum (CSS) in Alkanes by Stepwise Multiple Linear Regression (SMR) with Molecular Path Index Vector (VPM). *Chin. J. Magn. Reson.* **1999**, *16*, 243–254.
73. Mao, J.; Jain, A.K.; Jain, A. Artificial neural networks for feature-extraction and multivariate data projection. *IEEE Trans. Neural Netw.* **1995**, *6*, 296–317.
74. Vapnik, V. *Statistical Learning Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1998.
75. Sheng, G.; Huang, K.; Chi, Z.; Ding, H.; Xing, Y.; Lu, P.; Wang, Y. Preparation of 3-diazoindolin-2-imines via cascade reaction between indoles and sulfonylazides and their extensions to 2,3-diaminoindoles and imidazo[4,5-b]indoles. *Org. Lett.* **2015**, *46*, 5096. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.