RESEARCH ARTICLE

# Predicting the Yield of Pd-catalyzed Buchwald-Hartwig Amination using Machine Learning with Extended Molecular Fingerprints and Selected Physical Parameters

WeiRen Zhao[a, b] and Yang Li*[a,b]

[a] State Key Laboratory of Fine Chemicals, Dalian University of Technology, Dalian 116024, China. E-mail: chyangli@dlut.edu.cn
[b] School of Chemical Engineering, Ocean and Life Science, Dalian University of Technology, Panjin 124221, China

**Abstract:** Machine learning has gained attention due to its ongoing advancements and diverse applications. Within the field of homogeneous catalysis, a prominent area of research in machine learning revolves around predicting reaction yield in Pd-catalyzed Buchwald-Hartwig amination reactions. This study sought to determine the optimal descriptors for representing the both structural and physical information associated with the reaction, particularly focusing on product details. To achieve this, we assessed the utilization of product extended molecular fingerprints (PEMF) and selected physical parameters (SPP). The utilization of a random forest model incorporating these descriptors yielded promising results in the prediction of reaction yields in Pd-catalyzed Buchwald-Hartwig amination reactions. The model achieved an impressive $R^2$ value of 0.943. Data preprocessing on PEMF and sorting preprocessing on physical parameters resulted in a significant reduction in data size to 259 bits PEMF + 2 SPPs per prediction, much less than the two previous random forest models which utilized 480 physical parameters and 21,073 bits molecular fingerprints. Although establishing definitive correlations between SPPs and reaction yield presented challenges, our findings indicate that the presence of heavier atoms in the aryl halides may have a beneficial impact within the examined Pd-catalyzed Buchwald-Hartwig amination reactions, as compared to their analogues.

## Introduction

In recent years, machine learning (ML) has gained significant attention for its continued advancements and widespread applications across fields such as materials science and drug design.[1-2] ML has also proven valuable in the domain of homogeneous catalysis, enabling the prediction of reaction yield and selectivity, catalyst design, study of reaction mechanisms, optimization of reaction conditions, and automated experimentation.[3-5] Compared to traditional experimental methods, ML applications have the potential to improve productivity and cost-effectiveness in catalyst development, thus advancing environmentally friendly catalytic technologies.[6-7]

One of the significant research areas in ML for homogeneous catalysis is the prediction of reaction yield in Pd-catalyzed Buchwald-Hartwig amination reactions, which are known for their efficient formation of C-N bonds. Doyle et al. made contributions to this research area through their implementation of high-throughput experiments.[8] They also successfully predicted the yield using a random forest (RF) model, employing 480 physical parameters derived from DFT calculation, resulting in an $R^2$ value of 0.92 (Figure 1a). Glorius and colleagues, however, developed an RF prediction model using multiple fingerprint features (MFFs) as descriptors with 21,073 bits (Figure 1b), achieving a similar $R^2$ value as Doyle's model.[9] These MFFs are derived from molecular fingerprint (MF) based on the molecular structure of the reactants in the Pd-catalyzed Buchwald-Hartwig reaction. More advanced models, such as Bidirectional Encoder Representations from Transformers (BERT) in Figure 1c and Graph Neural Networks (GNNs) in Figure 1d, have shown superior predictive performance compared to the RF model.[10-14] However, these models come with increased complexity and computational requirements. Therefore, it is crucial to explore strategies that can enhance predictive accuracy, improve reaction interpretability, and reduce computational costs.

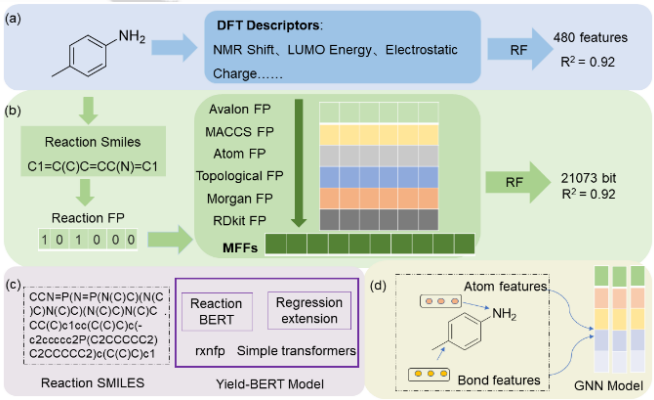



**Figure 1.** Previous ML models on predicting the yield of Pd-catalysed Buchwald-Hartwig amination: (a) Random Forest model using the DFT descriptors. (b) Random Forest model using MFFs as descriptors. (c) BERT model using SMILES as input. (d) GNN model using manually screened parameters as descriptors.

In this study, we aimed to develop a ML model for predicting the reaction yield of Pd-catalyzed Buchwald-Hartwig amination reactions, utilizing the same high-throughput experiments results from Doyle. It is noteworthy that certain criticisms have been leveled against the dataset, asserting its potential shortcomings. However, the original authors subsequently rebutted these claims, asserting the dataset's integrity. Consequently, a multitude of studies based on this dataset have proceeded under the assumption of its validity, directly employing it in their respective

investigations without further ado.[15-16] Several state-of-the-art methods were successfully developed and implemented using the same dataset, and a certain level of interpretability was achieved for the models. [17-20] Inspired by previous research, our approach integrates a RF model with novel descriptors, aiming to harness the strengths of both molecular fingerprints and physical parameters to effectively convey the structural and physical information associated with the reaction. We hope our model and new descriptors could improve the accuracy of reaction yield prediction while keeping computational costs low, and to provide insights into the mechanisms of Pd-catalyzed Buchwald-Hartwig amination reactions.

## Results and Discussion

Figure 2 outlines our approach for obtaining descriptors. It is worth noting that predicting reactions yield presents significant challenges due to the influence of reaction variables under investigation and potential side reactions.[21] To address this, we incorporated products information into our descriptors. We used the simplified molecular input line entry system (SMILES) to represent each reaction, including aryl halides, bases, ligands, additives, and products.[22] These SMILES were then converted to a product extended MF (PEMF) to capture structural details, particularly products information. To alleviate the computational burden associated with DFT calculation on physical parameters, we employed a sorting process on the physical parameters generated by Mordred, aiming to identify the most important selected physical parameters (SPP) for optimizing manual parameter screening.[23]

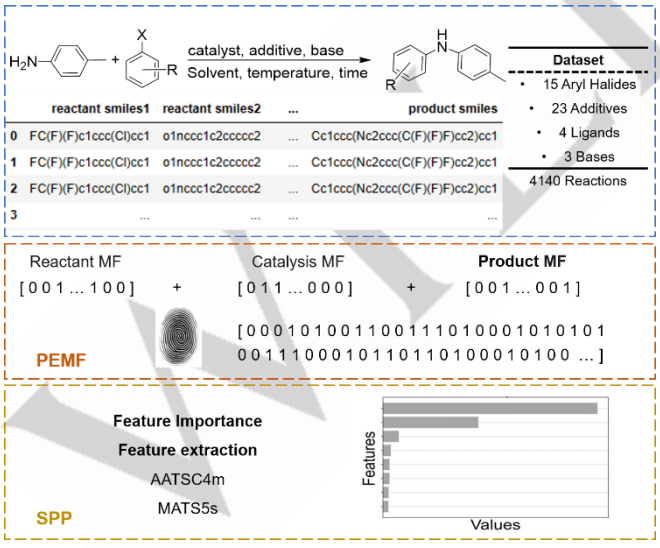



**Figure 2.** Descriptor generation in this work containing product extension molecular fingerprint and selected physical parameters.

As depicted in the workflow outlined in Figure 2, we employed Rdkit to generate product extended SMILES representations for each reaction, and subsequently converted those to PEMFs using six different molecular fingerprint types (MACCS, Topological, Atom, Rdkit, Morgan, and Avalon, shown in Table S1-S27 in Supporting Information).[24-27] Data preprocessing was applied to reduce the data size by eliminating duplicate values in the PEMF data, specifically by excluding columns that contained either zero or one for all reactions, resulting in a reduction from 2560 bits to 259 bits for each prediction.

Figure 3a demonstrates the PEMF descriptors (purple bar) outperformed the corresponding reactant-based MF descriptors (green bar) across all six fingerprint types, with $R^2$ values ranging from 0.918 to 0.939 (see Figure S1-S12), thereby demonstrating the significance of incorporating product information. The Avalon PEMF exhibited the highest performance with an $R^2$ value of 0.9386, surpassing the previously used MFFS descriptors, thereby justifying its selection for further analysis.

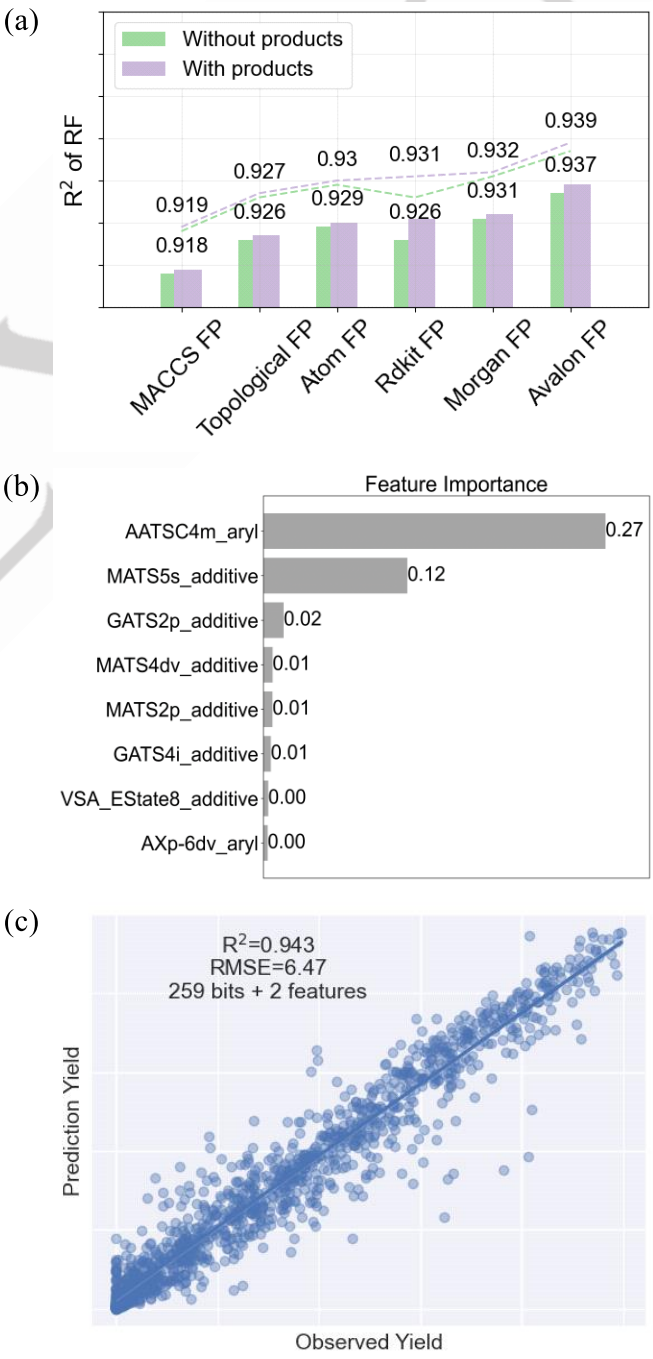
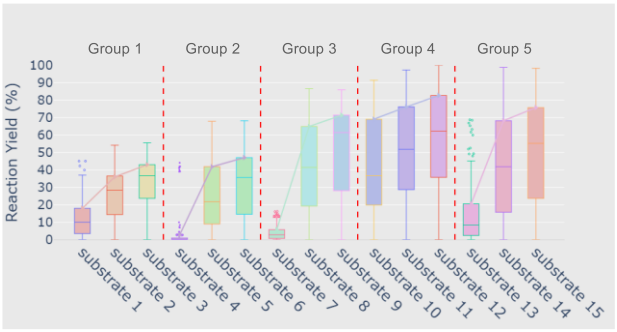
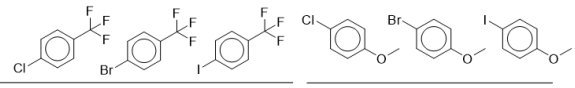


**Figure 3.** (a) The $R^2$ value for RF model using six type molecular fingerprints with and without the products information. (b) The significance of top eight physical parameters. (c) A superior RF model of this work was established with $R^2$ of 0.943 by utilizing the fewest features (259 bits + 2 features).

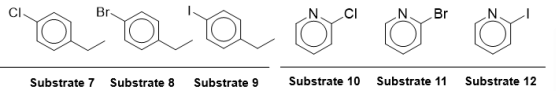
(a)



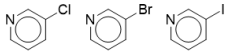

(b)



| | Substrate 1 | Substrate 2 | Substrate 3 | Substrate 4 | Substrate 5 | Substrate 6 |
|---|---|---|---|---|---|---|
| AATSC4m | -30.575 | -73.592 | -135.789 | -26.233 | -60.856 | -117.306 |
| Yield% (Q3) | 18.019 | 36.510 | 43.008 | 2.579 | 41.852 | 47.085 |



| | Substrate 7 | Substrate 8 | Substrate 9 | Substrate 10 | Substrate 11 | Substrate 12 |
|---|---|---|---|---|---|---|
| AATSC4m | -14.788 | -31.154 | -59.070 | -37.073 | -161.009 | -386.774 |
| Yield% (Q3) | 5.689 | 64.901 | 71.428 | 69.082 | 76.166 | 82.716 |



| | Substrate 13 | Substrate 14 | Substrate 15 |
|---|---|---|---|
| AATSC4m | -1.429 | -43.932 | -140.182 |
| Yield% (Q3) | 20.496 | 68.188 | 75.750 |

**Figure 4.** (a) Box plots illustrating the distribution of reaction yield associated with substrates. (b) The calculated AATSC4m values and reaction yield associated with substrates.

To capture the physical parameters of the reactants, the Mordred package was utilized to get physical parameters, based on the product extend SMILES representations. A RF regression model was employed to evaluate the significance of the total 4999 physical parameters, and the top eight influential features were identified (Figure 3b). The AATSC4m descriptor[28-30] of aryl halide substrates and the MATS5s descriptor[31-32] of additives showed the highest levels of significance at 27% and 12%, respectively. The significance sharply declined beyond the third parameter, dropping from 2% to 0%, leading to the selection of only the top two physical parameters as SPP descriptors. The histograms depicting the distribution of descriptors were plotted (Section 2, Figure S13), which reveal a broad coverage range and a balanced number of descriptors for each reactant, devoid of any instance where the number of descriptors for one reactant significantly surpasses that of others. Besides, two supplementary feature selection approaches, Lasso and Ridge, were also adopted to scrutinize all descriptors (see Section 2, Figures S14-S15, in SI). The divergent rankings of descriptor importance across these different methods underscore the diversity and meaningfulness of the descriptors selected for our analysis.

To ensure the reliability and validity of model evaluation results, a 10-fold cross-validation was adopted to assess the model's generalization capability ( see Table S28-S30 in SI). A RF model using all 4999 calculated physical parameters achieved an $R^2$ value of 0.936 (Figure S16 in Supporting Information), while a noticeable decline in the predictive performance was observed when using only two specific SPPs, with an $R^2$ value of 0.558 (Figure S17). However, a RF model using the combination of Avalon PEMF and SPPs achieved an improved $R^2$ value of 0.943 (Figure 3c, more details are shown in Table S28), surpassing the previous model reported by Doyle and Glorius with an $R^2$ value of 0.92. Importantly, the use of Avalon PEMF+SPP descriptors significantly reduced the dataset size to 259 bits MF and 2 physicochemical parameters, thereby improving its overall efficiency.

This data set is composed of a diverse array of chemical components, including 15 distinct aryl halides, 23 unique additives, 4 different ligands, and 3 separate bases (Figure 2). Each of these elements contributes to the formation of a total of 4140 reactions, each with its own corresponding yield value. To investigate the correlation between SPP (the AATSC4m descriptor of aryl halide substrates and the MATS5s descriptor of additives) with reaction yield in Pd-catalyzed Buchwald-Hartwig amination reactions, we analyzed the yield distributions for each aryl halide substrate and additive. For each aryl halide substrate, approximately 276 reactions were conducted, while each additive was associated with about 180 reactions. Figures 4a represented the resulting yield distributions of 15 aryl halide substrate using box plots. Figure 4b displayed the calculated AATSC4m values for aryl halide substrates. The reaction yield distributions by the Q3 value, which signified the third quartile of the data set, were also represented in Figure 4b.

The AATSC4m values consistently demonstrated negativity across all substrates, with large variation ranging from -1.492 to -386.774 among different chemical structures. At first glance, there appeared to be no significant correlations between the AATSC4m values and reaction yield. For instance, substrate 13 had the highest AATSC4m value of -1.492 but yielded poorly, with a Q3 value of 20.496 %, which still surpassed the yield of substrates 1, 4, and 7. Conversely, substrate 12 had the lowest AATSC4m value of -386.774 exhibited the highest yield with a Q3 value of 82.716%.

However, a closer examination revealed a clear trend. In Figure 4a, the yield distributions could be categorized into five distinct groups (Group 1-5), each containing three substrates. Within each group, there was a consistent increase in yield from low to high. Similarly, when observing the 15 substrates in Figure 4b, the same five groups were discernible, with higher structural similarity but varying halogen substituents (Cl, Br, and I) were observed. In each group, as the mass of the halogen substituent increased, the substrates exhibited more negative AATSC4m values and higher yield. For instance, in Group 1 consisting of substrates 1, 2, and 3 (4-chlorobenzotrifluoride, 4-bromobenzotrifluoride, and 4-iodobenzotrifluoride), the progressively negative AATSC4m values (-30.575, -73.592, and -135.789, respectively) correlated with higher yield, as demonstrated by the Q3 values (18.091%, 36.510%, and 43.008%, respectively) with increasing mass of the

halogen substituent. It is worth noting that comparing values across different groups did not follow the observed trend. For instance, substrate 3 in Group 1 has a more negative value of -135.789 compared to substrate 6 in Group 2, which has a value of -117.306. However, these two substrates exhibit comparable yield with Q3 values of 43.008% and 47.085%.

To elucidate this intriguing phenomenon, we conducted an in-depth examination of the interpretation of the AATSC4m descriptor, which represents a 2D topological descriptor closely associated with the averaged and centered Moreau-Broto autocorrelation of lag 4, weighted by mass (details refer to Supporting Information). Upon interpretation of the AATSC4m descriptor, it becomes apparent that its utilization for analyzing the correlation between aryl halide substrates and reaction yield presents challenges due to the diverse structures of individual components within aryl halides, thereby complicating the comprehensive assessment of their collective contributions. Nonetheless, when the structures of aryl halides exhibit greater similarity and the contributions of other components to the total AATSC4m values remain constant, it becomes feasible to discern the impact of specific substituents. Moreover, these findings support our observations in Figure 4, indicating a clear trend between the AATSC4m values and reaction yield within the same group, although this trend is not consistent when comparing values across different groups.

Despite these challenges, these findings provide valuable insights into the association between aryl halide substrates and reaction yield, particularly in elucidating the influence of specific substituents in the presence of closely related structures. Within the context of the Pd-catalyzed Buchwald-Hartwig amination reactions examined in this work, the SPP demonstrate that aryl halides containing heavier atoms compared to their analogous structures exhibit more negative AATSC4m values, subsequently leading to higher reaction yield. This observed correlation provides a mechanistic insight that the inclusion of heavier atoms in the aryl halides may have a beneficial impact on the reaction yield of the Pd-catalyzed Buchwald-Hartwig amination reactions under investigation.

After evaluating the AATSC4m descriptor, the relationship between MATS5s descriptors of 23 additives and the reaction yield in Pd-catalyzed Buchwald-Hartwig amination reactions were investigated. The structures of all the additives can be found in Figure S18, and the distribution of reaction yield associated with these additives is shown in Figure S19. Detailed information regarding the calculated MATS5s values, corresponding reaction yields with respective Q3 values can be found in Table S31.
The MATS5s descriptors ranged from -0.612 to 0.794, while the reaction yield with a Q3 value varied from 14.850% to 76.635%. Despite conducting a thorough analysis, we encountered difficulties in establishing definitive correlations between the MATS5s values of the additives and the resulting reaction yield. Further examination of the MATS5s descriptor, which is a 2D topological descriptor closely related to atomic electronegativity within the structure (details refer to Supporting Information), revealed the structural complexity that hindered our comprehensive assessment of their collective contributions to the reaction yield.

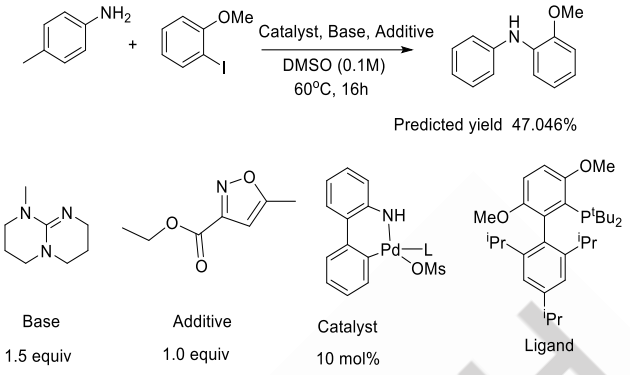



**Figure 5.** Predicted yield for a designed reaction

In an attempt to perform an inverse analysis of the prediction model, we designed a reaction that was initially excluded from the HTE dataset for this purpose, as depicted in Figure 5. By utilizing the structural template of **substrate 6**, we introduced 2-Iodoanisole - a compound featuring a less negative AATSC4m value of -132.278 - in conjunction with 4-methylaniline for a Buchwald-Hartwig amination. Consistent with our prior research findings, the reaction incorporating 2-Iodoanisole was anticipated to exhibit superior efficacy, with a projected yield of 47.046%, compared to reactions utilizing **substrate 6**, which yielded 45.982% under identical conditions and possesses an AATSC4m value of -117.306.

## Conclusion

The Pd-catalyzed Buchwald-Hartwig amination reactions, known for their efficient formation of C-N bonds, have attracted considerable attention in machine learning research. In order to develop machine learning strategies aimed at enhancing predictive accuracy, improving reaction interpretability, and reducing computational costs, we developed two descriptors, namely PEMF and SPP.

To determine effective descriptor for representing the structural information of the Pd-catalyzed Buchwald-Hartwig amination reaction, six types of molecular fingerprints, with and without product information, were evaluated. The Avalon PEMF parameter exhibited the highest performance with $R^2$ value of 0.939, thereby demonstrating the significance of incorporating product information.

To identify appropriate descriptors for representing the physical information, a sorting analysis of 4999 physical parameters was conducted, resulting in the selection of two SPP descriptors.

Employing all 4999 physical parameters yielded an $R^2$ value of 0.936, whereas using only the SPP parameters led to a noticeable decrease in predictive performance with $R^2$ value of 0.558. However, when combined the Avalon PEMF with SPP, the $R^2$ value improved to 0.943, surpassing the performance of previous RF models and enhancing overall efficiency.

Furthermore, preprocessing PEMF data and sorting physical parameters reduced the dataset size to 259 bits PEMF + 2 physical parameters per prediction, which was advantageous for our approach.

Although establishing direct correlations between these SPPs and reaction yield was challenging, our findings suggest that the presence of heavier atoms in aryl halides may have a positive impact on Pd-catalyzed Buchwald-Hartwig amination reactions compared to their analogues. Unfortunately, we encountered difficulties in establishing concrete correlations between MATS5s values of additives and resulting reaction yield. Then the study devised a new reaction that was not originally part of the HTE dataset, for an inverse analysis.

## Acknowledgements

This research was supported by the National Science Foundation of China (21903010).

**Keywords:** Buchwald-Hartwig • Random Forest • Machine Learning

## Conflict of Interests

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article. The source code and data sets are also available at https://github.com/ZWR0/PEMF-SPP/tree/main.

[1] K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh, J. Zou, *Cell.* **2023**, *186*, 1772–1791.

[2] P. Kalhor, N. Jung, S. Bräse, C. Wöll, M. Tsotsalas, P. Friederich, *arXiv preprint* **2023**, DOI: 10.1002/adfm.202302630.

[3] S. Gallarati, R. Fabregat, R. Laplaza, S. Bhattacharjee, M. D. Wodrich, C. Corminboeuf, *Chem. Sci.* **2021**, *12*, 6879–6889.

[4] N. I. Rinehart, R. K. Saunthwal, J. Wellauer, A. F. Zahrt, L. Schlemper, A. S. Shved, R. Bigler, S. Fantasia, S. E. Denmark, *Science.* **2023**, *381*, 965–972.

[5] S. Shilpa, G. Kashyap, R. B. Sunoj, *J. Phys. Chem. A.* **2023**, *127*, 8253–8271.

[6] G. Reginato, P. Sadler, R. D. Wilkes, *Org. Process Res. Dev.* **2011**, *15*, 1396–1405.

[7] S. Nath, E. Yadav, A. Raghuvanshi, A. K. Singh, *Catal. Sci. Technol.* **2023**, *13*, 7085–7099.

[8] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science.* **2018**, *360*, 186–190.

[9] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, *Chem.* **2020**, *6*, 1379–1390.

[10] Y. Li, C. Y. Hsieh, R. Lu, X. Gong, X. Wang, P. Li, S. Liu, Y. Tian, D. Jiang, J. Yan, Q. Bai, H. Liu, S. Zhang, X. Yao, *Nat. Mach. Intell.* **2022**, *4*, 645–651.

[11] E. M. Collins, K. Raghavachari, *J. Phys. Chem. A.* **2021**, *125*, 6872–6880.

[12] E. M. Collins, K. Raghavachari, *J. Chem. Theory Comput.* **2023**, *19*, 2804–2810.

[13] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, *arXiv preprint* **2019**, DOI: 10.18653/v1/N19-1423.

[14] S. W. Li, L. C. Xu, C. Zhang, S. Q. Zhang, X. Hong, *Nat. Commun.* **2023**, *14*, 3569–3579.

[15] K. V. Chuang, M. J. Keiser, *Science*. **2018**, *362*(6416): eaat8603.

[16] J. G. Estrada, D. T. Ahneman, R. P. Sheridan, *Science*. **2018**, *362*(6416): eaat8763.

[17] P. Schwaller, A. C. Vaucher, T. Laino, *Mach. learn.: sci. technol*. **2021**, *2*, 015016.

[18] S. Singh, R. B. Sunoj, *Digital Discovery*. **2022**, *1*, 303-312.

[19] LY. Chen, YP. Li, *J. Cheminform*. **2024**, *16*, 74.

[20] S. Laverty, S. Dey, A. F. Zahrt, *Chem.* **2024**, *10*, 1623-1626.

[21] M. Saebi, B. Nan, J. E. Herr, J. Wahlers, Z. Guo, A. M. Zurański, T. Kogej, P. O. Norrby, A. G. Doyle, N. V. Chawla, O. Wiest, *Chem. Sci.* **2023**, *14*, 4997–5005.

[22] D. Weininger, *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[23] H. Moriwaki, Y. S. Tian, N. Kawashita, T. Takagi, *J. Cheminformatics.* **2018**, *10*, 4-15.

[24] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.

[25] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[26] R. Nilakantan, N. Bauman, J. S. Dixon, *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.

[27] Rdkit: Open-source chemoinformatics and machine learning. http://www.rdkit.org.

[28] G. Moreau, P. Broto, *Nouv. J. Chim.* **1980**, *4*, 359-360.

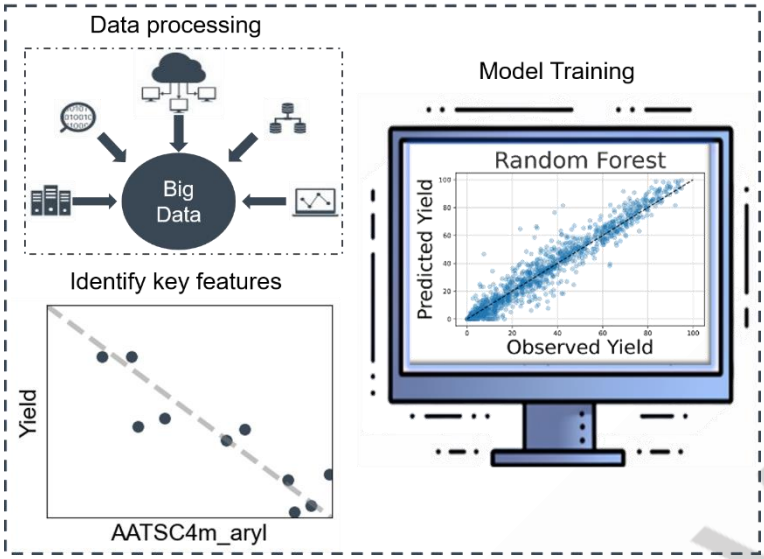
[29] B. Hollas, *J. Math. Chem.* **2003**, *33*, 91–101.

[30] M. Wagener, J. Sadowski, *J. Am. Chem. Soc.* **1995**, *117*, 7769–7775.

[31] Q. Huang, W. Song, L. Wang, *Chemosphere.* **1997**, *35*, 2849–2855.

[32] B. K. Lemont, H. H. Lowell, *J. Pharm. Sci.* **1986**, *76*, 269–270.

## Entry for the Table of Contents




This work focuses on balancing computation cost and model accuracy through Machine Learning for forecasting the reaction yield in Pd-catalysed Buchwald-Hartwig amination reactions. The study evaluated the effectiveness of product extended molecular fingerprints (PEMF) and selected physical parameters (SPP) as critical features. The model achieved an $R^2$ value of 0.943 using a smaller dataset. Further exploration of the reaction mechanisms reveals that the presence of heavier atoms in the aryl halides may have a beneficial impact within the examined Pd-catalysed Buchwald-Hartwig amination reactions.