ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Design of Experimental Conditions with Machine Learning for Collaborative Organic Synthesis Reactions Using Transition-Metal Catalysts

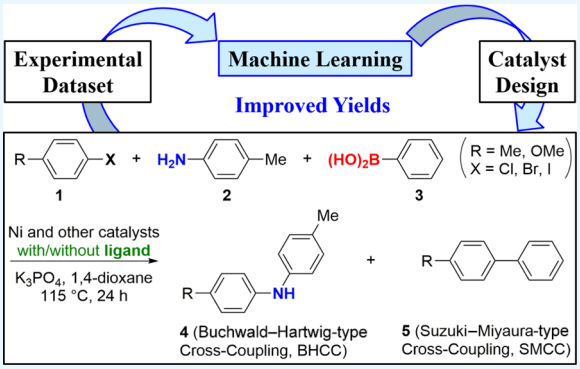
Tomoya Ebi, Abhijit Sen, Raghu N. Dhital, Yoichi M. A. Yamada, and Hiromasa Kaneko*

Cite This: *ACS Omega* 2021, 6, 27578−27586

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: To improve product yields in synthetic reactions, it is important to use appropriate catalysts. In this study, we used machine learning to design catalysts for a reaction system in which both Buchwald−Hartwig-type and Suzuki−Miyaura-type cross-coupling reactions proceed simultaneously. First, using an existing dataset, yield prediction models were constructed with machine learning between experimental conditions, including the substrate and catalyst and the yields of the two products. Seven methods for calculating both the substrate and catalyst descriptors were proposed, and the predictive ability of the yield prediction models was discussed in terms of the descriptors and machine learning methods. Then, the constructed models were used to predict the compound yields for new combinations of substrates and catalysts, and the predictions were experimentally validated with high reproducibility, confirming that machine learning can predict yields from experimental conditions with high accuracy. In addition, to design catalysts that will improve the yields in our dataset, we added datasets collected from scientific papers and designed catalyst ligands. The proposed catalyst candidates were tested in actual synthetic experiments, and the experimental results exceeded the existing yields.



INTRODUCTION

With the shortening of product life cycles, the development of catalysts to promote chemical reactions is becoming important for manufacturing highly functional chemical materials such as pharmaceuticals and electronic materials in short time frames. In general catalyst development, catalysts are designed based on the knowledge and experience of experimental scientists, who also refer to prior examples and technologies described in scientific papers and patents. The designed catalysts are then synthesized and used in a target chemical reaction, and the experimental results are checked. The catalysts are then redesigned after feedback of the results. This cycle is repeated to develop a target catalyst with a high reaction performance while discussing the design strategy. Importantly, this cycle of catalyst design, synthesis, and activity evaluation/verification can be both time- and cost-intensive, which increases the development period.

With the improved performance of computers and molecular simulations, it has become possible to analyze reaction mechanisms and catalytic reaction intermediates using computational science. However, accurate molecular simulations are computationally time-consuming and thus unsuitable for preprocessing and simulating massive numbers of catalyst candidates to search for a promising one. In this study, we focused on using machine learning to accelerate catalyst design and development. In the quantitative structure−activity relationship[1,2] and quantitative structure−property relationship[3,4] models, a mathematical function $y = f(x)$ is constructed between activities and properties y and molecular descriptors x using a compound dataset and machine learning. Some examples of machine learning methods used for model construction are partial least squares (PLS),[5] ridge regression (RR),[6] least absolute shrinkage and selection operator (LASSO),[6] elastic net (EN),[6] support vector regression (SVR),[7] Gaussian process regression (GPR),[8] random forest (RF),[9] gradient-boosting decision tree (GBDT),[10] extreme gradient boosting (XGB),[11] and light gradient-boosting (LGB)[12−14] models. The prediction performance of each method depends on the compound dataset, and trial and error are required to select an appropriate method for a given dataset. Physicochemical properties and pharmacological activities have been predicted in the areas of chemistry and drug discovery using machine learning.[15,16]

Received: September 2, 2021
Accepted: September 28, 2021
Published: October 5, 2021

When y is catalytic activity, machine learning models can be used to estimate catalytic activities from catalyst chemical structures without their synthesis. Even when the detailed mechanism of the catalytic activity within the organic reaction is unknown, if we can propose features x that are important in explaining the catalytic activity, we can obtain the correlation between x and y using an experimental dataset and machine learning.

Yada et al. used LASSO to predict yields for the use of different tungsten catalysts in the epoxidation of alkenes with hydrogen peroxide as a model reaction.[17] Ahneman et al. predicted the reaction performance of various potentially inhibitory ligands for the palladium-catalyzed Buchwald−Hartwig cross-coupling of aryl halides with 4-methylaniline using machine learning methods.[18] While machine learning has been applied to single organic synthetic reactions, it has not been applied to multiple reactions proceeding simultaneously. Furthermore, in actual synthetic experiments, not only the catalyst but also experimental conditions such as the reaction temperature, substrate, bases, and ligands can change. In such experimental datasets, experimental conditions other than the catalyst are also important and must be considered in machine learning.

The objective in this study was to construct predictive models that consider multiple experimental conditions for multiple reactions proceeding competitively and then to search for experimental conditions with high catalytic activities using the constructed models. We considered not only the chemical structures of the catalysts but also the chemical structures of the substrates and other information as x and constructed models between x and yields y using machine learning. Since the appropriate way to describe structures of ligands, active metal centers, and reactants is unclear, we propose various methods (methods A, B, C, D, E, and F), including computing descriptors for each of them separately, calculating descriptors for assumed reaction intermediates, and adopting simplified descriptors. Subsequently, we used the model to search for promising experimental conditions such as catalysts and substrates to improve the conventional yields. In addition, we proposed a method to search for new experimental conditions by utilizing not only a target experimental dataset but also datasets collected from scientific papers. To validate the effectiveness of the proposed method, we verified the proposed experimental conditions by conducting actual syntheses.

METHODS

After the target reaction system in this study is described, methods A through F, which are the descriptor calculation methods developed in this study, are explained.

Target Reaction System. In this study, we used an experimental dataset for the reactions of aryl halides **1** with *p*-toluidine **2** (3.0 mol equiv) and phenylboronic acid **3** (1.5 mol equiv) using nickel catalysts (0.5 mol %) with/without ligands in the presence of K_3PO_4 (3 mol equiv), as shown in Figure 1. The reaction system includes two reaction pathways that occur simultaneously: Buchwald−Hartwig-type cross-coupling (BHCC)[19] and Suzuki−Miyaura-type cross-coupling (SMCC).[20] The SMCC reaction mainly proceeds when substrate **1** is an aryl chloride (X = Cl), whereas the BHCC reaction mainly proceeds when substrate **1** is an aryl iodide (X = I); however, the detailed mechanisms underlying this selectivity are not clear.[21] Therefore, we analyzed an

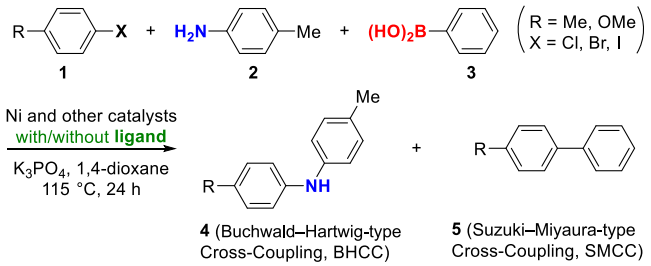


Figure 1. Target reaction system.

experimental dataset consisting of the yields of the products of the BHCC and SMCC reactions when the functional group X of substrate **1** is changed to Cl, Br, and I, the functional group R of substrate **1** is changed to Me and OMe, and the transition-metal catalyst is changed. All experimental conditions were the same except for substrate **1** and the transition-metal catalyst. The objective variables y were the yields [%] of compound **4** (product of the BHCC reaction) and compound **5** (product of the SMCC reaction). There were 73 samples in our dataset.

Method A. In method A, the molecular descriptors are calculated from the chemical structures of substrate **1** and the catalyst, and the calculated descriptors are combined to form x. The 200 molecular descriptors for substrate **1** and those for the catalyst are calculated with RDKit.[22] The RDKit descriptors contain basic descriptors such as the number of atoms for each atom type and molecular weight, descriptors including fragment information, topological descriptors, and physicochemical descriptors.[23] When there are other chemical structures such as ligands, their molecular descriptors are also calculated and combined into x.

Method B. In method A, the descriptors of the two compounds (substrate and catalyst) are simply combined, resulting in a large x and thus the risk of overfitting. In method B, x is reduced by assuming the reaction intermediate shown in Figure 2. Since the BHCC and SMCC reactions proceed along

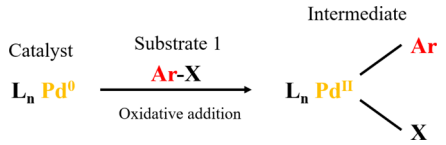


Figure 2. Assumed reaction intermediate.

a common reaction pathway until oxidative addition, a reaction intermediate that is expected to form in the common reaction pathway is prepared. Molecular descriptors are calculated for the reaction intermediate with RDKit and denoted as x, allowing us to consider information for substrate **1** and the catalyst simultaneously.

Method C. The size of x can be reduced from those of methods A and B by simplifying the information for the substrate and catalyst. In method C, the substrate functional group R (=Me, OMe), halogen X (=Cl, Br, I), the catalyst transition metal M (=Pd, Ni, Fe, etc.), and ligand L (=dppf, 1,5-cyclooctadiene, etc.) are represented by dummy variables of 0 (absence) or 1 (presence). If substituents and compounds are used, the variable is set to 1 and if not, it is set to 0. In addition, the total number of ligands is also added to x.

Method D. In method C, when designing new substrates and catalysts, we can use only R, X, L, and M that exist in

https://doi.org/10.1021/acsomega.1c04826
ACS Omega 2021, 6, 27578−27586

Table 1. Calculated r_{DCV}^2 and MAE_{DCV} of the Best Models for Each Yield Using Our Dataset

	4 (BHCC)			5 (SMCC)		
	regression method	r_{DCV}^2	MAE_{DCV}	regression method	r_{DCV}^2	MAE_{DCV}
Method A	XGB	0.449	11.0	XGB	0.508	15.5
Method B	XGB	0.099	15.4	GP	0.276	19.6
Method C	XGB	0.783	7.4	LASSO	0.563	15.2
Method D	RF	0.713	8.5	LASSO	0.495	16.4
Method E	RF	0.574	10.2	RF	0.507	16.2
Method F	LASSO	0.530	11.6	EN	0.446	16.1

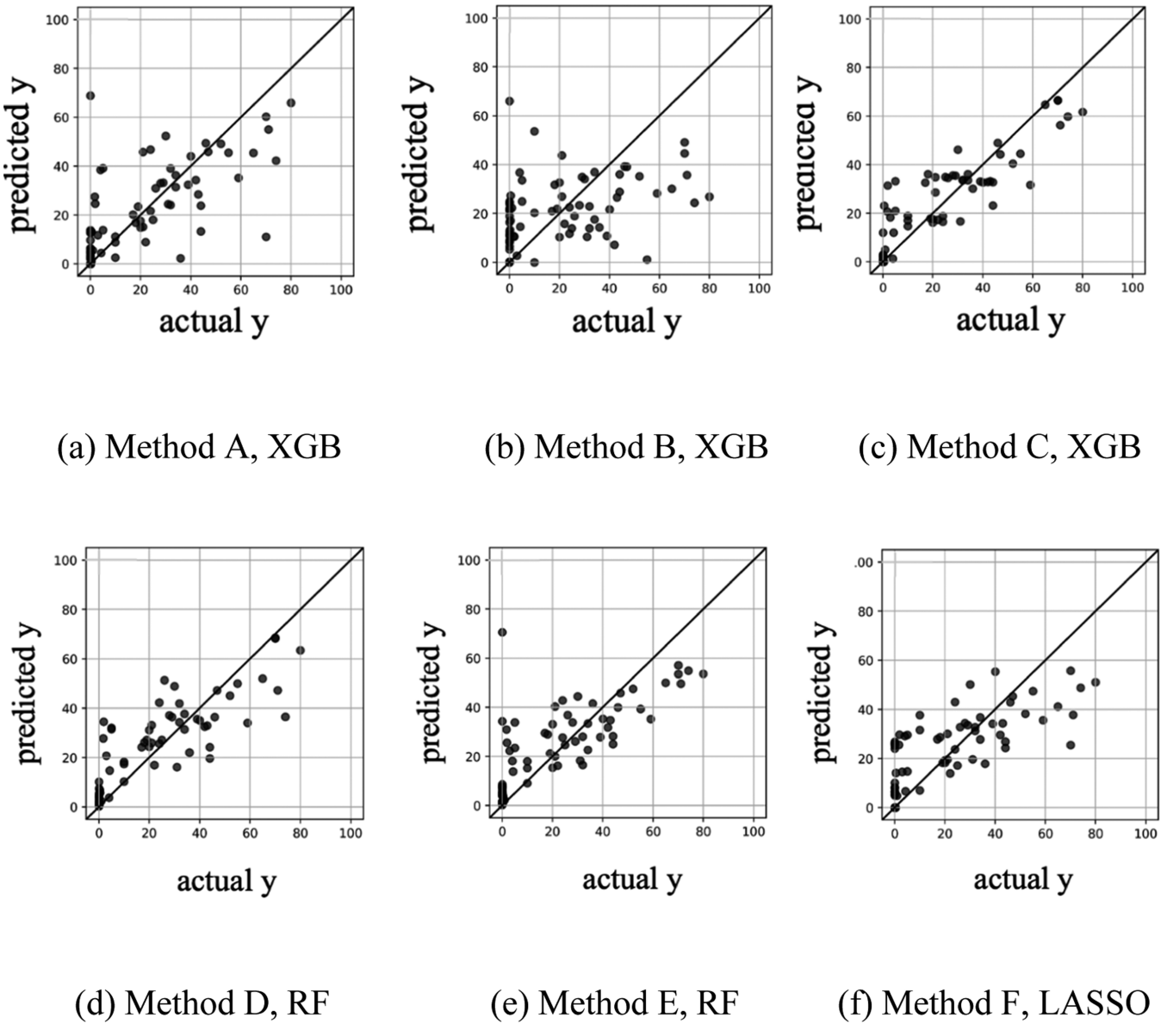


Figure 3. Actual vs predicted yields after DCV for each descriptor calculation method for compound 4 (BHCC).

previous substrates and catalysts from the employed dataset. In method D, the molecular descriptor of L in the catalyst is calculated to be able to design any ligands. RDKit is used to calculate the molecular descriptors. For catalysts with more than one L, the weighted average of each molecular descriptor relative to the coordination number is used as x. As a result, we can predict the y-values for new chemical structures and their combinations for L. The substrate and R, X, and M are represented by 0 and 1 as in method C.

Method E. While methods A–D employ the chemical structures of the molecules and simplified R, X, L, and M, electronic information for each molecule is also important for yield prediction. In method E, we use pymatgen,[24] an open-source Python library for materials analysis, to calculate the M descriptors and Spartan,[25] a molecular modeling software, to calculate the L descriptors for the transition-metal catalysts. In this study, five descriptors were calculated using pymatgen: atomic weight, atomic radius, number of outermost electrons,

https://doi.org/10.1021/acsomega.1c04826
ACS Omega 2021, 6, 27578–27586

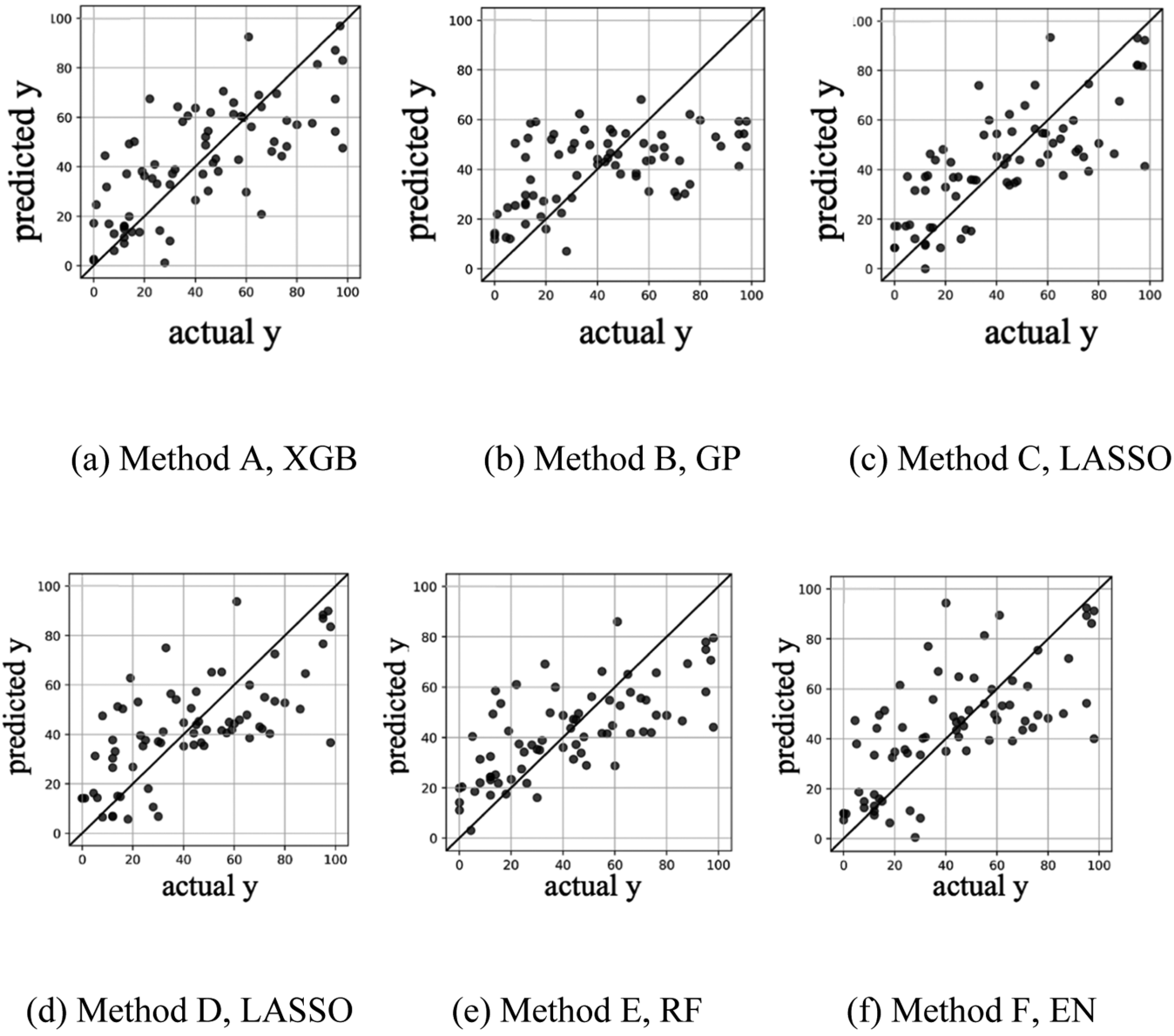


Figure 4. Actual vs predicted yields after DCV for each descriptor calculation method for compound **5** (SMCC).

volume, and electronegativity. Using Spartan, 18 descriptors such as molecular weight, polar surface area, HOMO, LUMO, polarizability, and number of hydrogen-bond acceptors and donors were calculated. When the catalysts have two or more L, the descriptors for each L are weighted by the coordination number and averaged as x. The substrate descriptors are calculated with RDKit.

Method F. If the chemical structures of the substrates and catalysts are similar in the two experiments, the yields will also be similar. In method F, the similarities in the chemical structures of the substrate and catalyst in the training data are combined into x. The similarity in this study is the Tanimoto coefficient calculated based on Morgan fingerprints in RDKit. When the number of samples in the training data is m, the number of descriptors is $2m$. When there are other chemical structures such as ligands, their structural similarities are added to x as descriptors.

Applicability Domain. Applicability domains (ADs)[26] for the yield prediction models were set since the performance of the model is unreliable when predicting new samples in extrapolation regions. In this study, we used the distances calculated by the k-nearest-neighbor algorithm[27] as an indicator of an AD. When a new sample is obtained, the average of the Euclidean distances of the k-nearest-neighbor samples from the new sample is calculated. When the average is lower than the threshold, the sample is inside the AD. The threshold was set as 99.6% of the averages in the ascending order in the training data. The threshold comes from the three-sigma rule,[28] and the probability of samples within the AD would be 0.996.

RESULTS AND DISCUSSION

Our dataset consisted of 73 experiments with different experimental conditions for the substrates and catalysts. The y-variables are the yields [%] of compound **4** (BHCC) and compound **5** (SMCC), as shown in Figure 1. Since the yield of compound **4** did not exceed 10% when X was Cl (aryl chloride) in substrate **1** in our dataset, we aimed to find experimental conditions with yields exceeding 10% using the aryl chloride. The proposed methods A−F were used to

https://doi.org/10.1021/acsomega.1c04826
ACS Omega 2021, 6, 27578−27586

investigate x. For each x, PLS, LASSO, RR, EN, SVR, GPR, RF, GBDT, XGB, and LGB were used to construct regression models. Because of the small sample size, we used double cross-validation (DCV)[29] to evaluate the yield predictive ability of the models, where the outer cross-validation (CV) was a leave-one-out CV and the inner CV was a fivefold CV. Since the outer CV was a leave-one-out CV, there was only one way to separate training data and test data. In other words, we repeated one sample for test data and the rest for training data for the total number of samples. The hyperparameter in each regression method was optimized in the inner CV. Candidates for the number of components in PLS were 1, 2, ..., and 20, candidates for the regularization parameter in RR are 2^{-15}, 2^{-14}, ..., and 2^{9}, candidates for the regularization parameter in LASSO are 2^{-15}, 2^{-14}, ..., and 2^{-1}, candidates for the regularization parameter in EN are 2^{-15}, 2^{-14}, ..., and 2^{-1}, candidates for the weight in EN are 0, 0.01, 0.02, ..., and 1, candidates for the regularization parameter in SVR are 2^{-5}, 2^{-4}, ..., and 2^{10}, candidates for the tolerated error are 2^{-10}, 2^{-9}, ..., and 2^{0}, candidates for the parameter in the Gaussian kernel function are 2^{-20}, 2^{-19}, ..., and 2^{10}, and candidates for the rate of the number of X used in RF are 10, 20, ..., and 90%.

For each x, DCV was performed for all the regression analysis methods, and the coefficient of determination after DCV (r_{DCV}^2) was calculated. Then, the regression analysis method with the largest r_{DCV}^2 was selected for each x. Table 1 shows the prediction results of the regression methods selected for each x, where MAE_{DCV} indicates the mean absolute error after DCV. In addition, plots of the measured and predicted yields after DCV are shown for each x in Figure 3 for compound **4** (BHCC) and in Figure 4 for compound **5** (SMCC). As observed in Table 1, method C had the highest r_{DCV}^2 and lowest MAE_{DCV} for the yields of both compounds **4** and **5**, which indicated that method C had the highest prediction accuracy. As shown in Figures 3 and 4, there were no outliers in the results of method C, and the yields were well predicted from low to high values. This is because the simplified substrate and catalyst information can properly model the relationship between x and y, and the size of x is reduced to prevent overfitting. On the other hand, method B had the lowest r_{DCV}^2 and highest MAE_{DCV} for the yields of both compounds **4** and **5**, indicating that method B had the lowest prediction accuracy. This illustrated that it would be difficult to construct predictive models with descriptors for the assumed reaction intermediates shown in Figure 2. We confirmed that appropriate modeling methods can be discussed while simultaneously considering substrate and catalyst descriptors from synthetic experiments and regression analysis methods.

To verify the effectiveness of the constructed yield prediction models, the yields of compounds **4** and **5** were predicted from 63 experimental conditions including new combinations of substrates and catalysts that differed from the combinations in our existing dataset. The model constructed with method C and XGB was used to predict the yield of compound **4**, and the model constructed with method C and the LASSO was used to predict the yield of compound **5**. From the prediction results, we performed synthesis experiments using the experimental conditions with the highest predicted yields of compounds **4** and **5**, and the actual yields were tested using the same solvent, base, reaction time, and reaction temperature as those in our dataset. Table 2 shows the predicted and measured yields of compounds **4** and **5**. The difference between the predicted and measured yields of compounds **4** and **5** were low for both samples 1 and 2, indicating that the yield prediction models had a high accuracy. It was confirmed that the proposed method could construct yield prediction models with a high predictive ability for a system in which two reactions proceed simultaneously; furthermore, the models can accurately predict yields for new experimental combinations of substrates and catalysts with high reproducibility.

Table 2. Predicted and Experimental Yields for New Combinations of Substrates and Catalysts

		sample 1[a]	sample 2[b]
yield of **4** (BHCC) [%]	prediction	70	0
	experiment	71	0
yield of **5** (SMCC) [%]	prediction	14	94
	experiment	16	96

[a]R is OMe, X is Cl, and the catalyst is $Ni(acac)_2(tolNH_2)_2$. [b]R is Me, X is I, and the catalyst is $Pd(dppf)Cl_2$.

Although yield prediction models for compounds **4** and **5** were developed using only our dataset, and new experimental conditions were designed by changing the substrate and catalyst, we could not meet the target yield values. To address this, it was necessary to expand the ADs of the yield prediction models and design new experimental conditions. Therefore, after the two samples in Table 2 were added to the initial 73 samples, we collected experimental datasets described in scientific papers on BHCC and SMCC reactions[30−33] and added them to our dataset to increase the scope of the catalyst structures and experimental conditions. The added experimental datasets included catalysts with transition metals such as Ni, Pd, Cu, Fe, and so forth and ligands that are different from our dataset. The collected datasets also included data for ligands, which play an important role in stabilizing and activating the central metal atom of the catalyst and fine-tuning the reaction selectivity. While it is necessary to represent not only the substrates and catalysts but also the ligands as x, it is possible to design all three. For the BHCC reaction, 150 samples were added for a total of 225. For the SMCC reaction, 55 samples were added for a total of 130.

In this analysis, to explore a wide range of experimental conditions, we used methods A and F, which enabled abstract representations of x. For each method, the different regression analysis methods (PLS, LASSO, RR, EN, SVR, GPR, RF, GBDT, XGB, and LGB) were used to construct yield prediction models, and the predictive ability of the models was evaluated with DCV. The values of r_{DCV}^2 and MAE_{DCV} for the models with the best DCV results for each yield and each descriptor calculation method are shown in Table 3, and plots of the measured and predicted yields are shown in Figure 5. Although the models could predict the overall trends of the yields for both BHCC and SMCC, there existed samples in which the predicted values became constant in BHCC. However, this phenomenon occurred mainly at low yields and would not be a problem in models that predict high yields. The prediction errors of method F were lower than those of method A, especially at higher yields. This is a desirable result when designing experimental conditions for higher yields. Based on the prediction results, a combination of method F and GP was used to predict the yield of compound **4** (BHCC), and a combination of method F and LASSO was used to predict the yield of compound **5** (SMCC).

https://doi.org/10.1021/acsomega.1c04826
ACS Omega 2021, 6, 27578−27586

Table 3. Calculated r_{DCV}^2 and MAE_{DCV} of the Best Models for Each Yield Using Our Dataset and Datasets Collected from Scientific Papers

	4 (BHCC)			5 (SMCC)		
	regression method	r_{DCV}^2	MAE_{DCV}	regression method	r_{DCV}^2	MAE_{DCV}
Method A	XGB	0.449	11.0	XGB	0.508	15.5
Method F	LASSO	0.530	11.6	EN	0.446	16.1

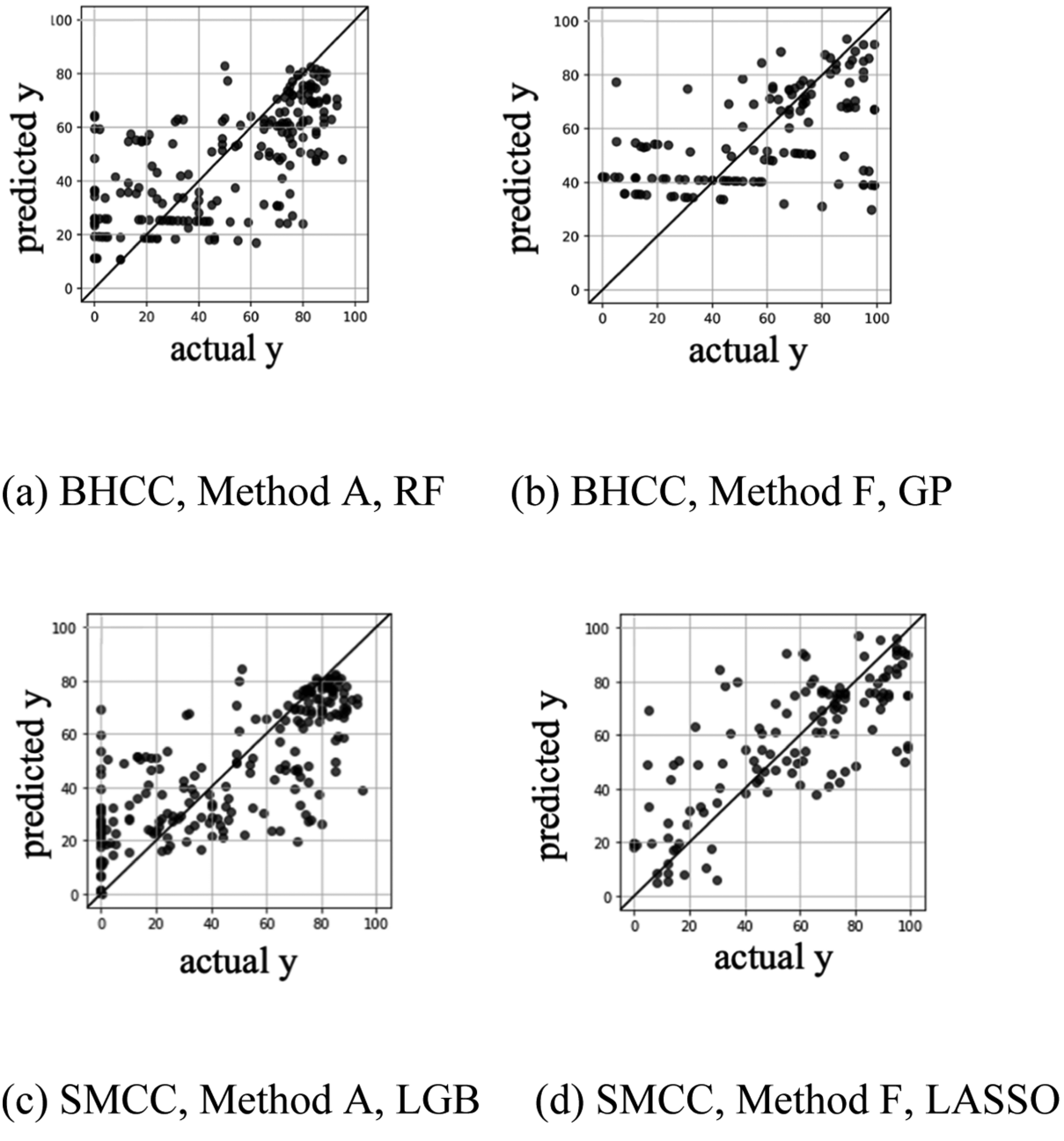


Figure 5. Actual vs predicted yields for each yield and the descriptor calculation method.

The model with the best DCV results for each yield was used to investigate new experimental conditions to achieve the target yield. In this study, we designed experimental conditions with higher predicted yields than existing combinations of substrates and catalysts by adding ligands. When the functional group X = Cl (aryl chloride) was used in substrate **1**, the yield of compound **4** did not exceed 10% in our dataset, and thus, we aimed to find experimental conditions that would surpass this. Since we focused on the combination of catalysts and ligands, we used the same experimental conditions for the solvent, base, reaction time, and reaction temperature as in our dataset. For the experimental catalysts and ligands, 33 catalysts from our dataset and 23 from collected papers[30−33] were employed, and 52,000 compounds from the Namiki Shoji database[34] were used as ligands.

A total of 1,716,759,000 new experimental conditions combining catalysts and ligands were input into the yield prediction models, and the experimental conditions with high

https://doi.org/10.1021/acsomega.1c04826
ACS Omega 2021, 6, 27578−27586

Table 4. Estimated and Experimental Yields for Each Experimental Condition

			yield of **4** (BHCC) [%]		yield of **5** (SMCC) [%]	
entry	catalyst	ligand	prediction	experiment	prediction	experiment
1	$Ni(acac)_2$	L2	36	6.3	28	15
2	$Ni(acac)_2$	L3	35	0	35	0
3	$Ni(acac)_2$	dppf	34	9.3	38	14
4	$Ni(acac)_2$	dppb	35	0	38	25
5	$Ni(acac)_2$	Xphos	35	33	37	30
7	$Ni(acac)_2(tolNH_2)_2$	Xphos	16	3.4	4.5	9.0
6	NiI_2	Xphos	15	4.4	51	80
8	$NiCl_2$	Xphos	15	7.8	35	37
9	$NiCl_2$	L2	19	0	39	20
10	$NiCl_2$ $(6H_2O)$	L2	19	0	44	42
11	$NiCl_2$ $(6H_2O)$	L4	17	0	28	7.8
12	$NiCl_2$ $(6H_2O)$	L1	17	0	44	28
13	$Ni(acac)_2$	Ruphos	15	12	16	35
14	$Ni(acac)_2$	Sphos	15	4.7	16	11
15	$Ni(acac)_2$	L5	12	5.1	16	23
16	$Ni(acac)_2(tolNH_2)_2$	L6	10	9.1	12	37
17	$Ni(acac)_2(tolNH_2)_2$	L7	7.4	6.4	13	31
18	$Ni(acac)_2(tolNH_2)_2$	L8	7.3	12	11	32
19	$Ni(acac)_2(tolNH_2)_2$	L9	7.1	15	10	16
20	$Ni(acac)_2(tolNH_2)_2$	L10	6.4	1.0	5.6	15

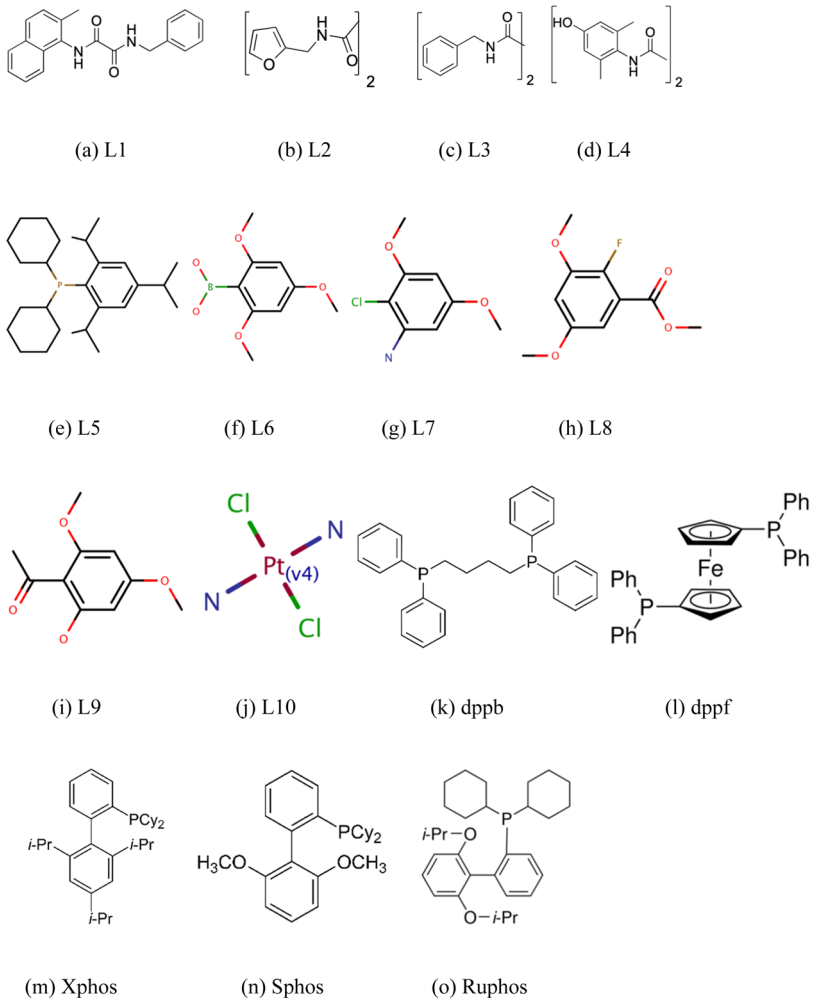


Figure 6. Ligand structures.

https://doi.org/10.1021/acsomega.1c04826
ACS Omega 2021, 6, 27578−27586

predicted yields were selected. The predicted and actual yields for the selected experimental conditions are shown in Table 4, and the ligand structures are shown in Figure 6. The MarvinView[35] software package, developed by ChemAxon, was used to visualize the chemical structures. Simplified molecular input line entry system (SMILES) of the ligandes is shown in Table S1 in Supporting Information. The prediction errors in Table 4 come from the errors in the training data, as shown in Figure 5. Furthermore, the reason why the prediction errors in Table 4 were larger than those in Table 2 would be that new catalysts were explored in Table 4, while the models predicted yields of new combinations of existing substrates and catalysts. For the aryl chloride substrate, when nickel(II) bis(acetylacetonate) ($Ni(acac)_2$) was used as the transition-metal catalyst and 2-dicyclohexylphosphino-2′,4′,6′-triisopropyl-1,1′-biphenyl (Xphos) was used as the ligand, the actual yield of compound **4** (BHCC) was 33%. This greatly exceeds the 10% yield of the BHCC reaction with $Ni(acac)_2$ as the catalyst without the use of a ligand. In addition, the yield of compound **4** exceeded 10% using not only the existing ligands but also compounds from the Namiki Shoji database that have not been used as ligands. These results confirmed that the proposed method can be used to find new experimental conditions that will exceed the existing yields.

CONCLUSIONS

In this study, we targeted a reaction system where BHCC and SMCC reactions proceed simultaneously and analyzed an experimental dataset with machine learning. We proposed six methods to represent the two reactions with the substrates and catalysts as descriptors and constructed yield prediction models using regression methods for each reaction. When only our dataset was used, method C, in which the chemical structures of the substrates and catalysts are simply represented as substituents, had the highest prediction accuracy. After predicting the yields for new combinations of substrates and catalysts using the constructed models, the predictions were confirmed to be in close agreement with the actual experimental results by synthetic experiments.

Then, datasets collected from scientific papers were added to our dataset, and potential combinations of catalysts and ligands were searched to exceed the existing yields with an aryl chloride substrate. Although there were no experimental data with ligands in our dataset, it was possible to search for ligands by adding collected datasets that included ligands as an experimental condition. A total of 1,716,759,000 new experimental conditions combining catalysts and ligands were input into the yield prediction models, and the experimental conditions with high predicted yields were selected. While the maximum yield in the existing data of the BHCC reaction was 10% with the aryl chloride substrate, it was experimentally confirmed that a yield of 33% could be achieved by adding the ligand predicted with the developed model.

Although it is important to discuss the mechanism that determines the selectivity of SMCC and BHCC based on the constructed models, this is difficult at present and, thus, is a future challenge. In summary, we have accurately predicted yields from experimental conditions, proposed experimental conditions exceeding the conventional yields, and experimentally verified the proposed reaction conditions. We expect the proposed method to accelerate catalyst development.

DATA AND SOFTWARE AVAILABILITY

Research data are not shared since experimental data are considered proprietary by the organization, and the algorithms of the proposed method are described in the article.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.1c04826.

SMILES of ligands in Table 4 (PDF)

AUTHOR INFORMATION

Corresponding Author

Hiromasa Kaneko − *Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan; RIKEN Center for Sustainable Resource Science, Wako, Saitama 351-0198, Japan;* orcid.org/0000-0001-8367-6476; Email: hkaneko@meiji.ac.jp

Authors

Tomoya Ebi − *Department of Applied Chemistry, School of Science and Technology, Meiji University, Kawasaki, Kanagawa 214-8571, Japan*

Abhijit Sen − *RIKEN Center for Sustainable Resource Science, Wako, Saitama 351-0198, Japan*

Raghu N. Dhital − *RIKEN Center for Sustainable Resource Science, Wako, Saitama 351-0198, Japan*

Yoichi M. A. Yamada − *RIKEN Center for Sustainable Resource Science, Wako, Saitama 351-0198, Japan;* orcid.org/0000-0002-3901-5926

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c04826

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by a Grant-in-Aid for Scientific Research (KAKENHI) (grant number JP19K15352) from the Japan Society for the Promotion of Science and Informatics Data Science Promotion Program from RIKEN CSRS.

ABBREVIATIONS

PLS, partial least squares; RR, ridge regression; LASSO, least absolute shrinkage and selection operator; EN, elastic net; SVR, support vector regression; GPR, Gaussian process regression; RF, random forest; GBDT, gradient-boosting decision tree; XGB, extreme gradient boosting; LGB, light gradient boosting; BHCC, Buchwald−Hartwig-type cross-coupling; SMCC, Suzuki−Miyaura-type cross-coupling; AD, applicability domain; MAE, mean absolute error

REFERENCES

(1) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.

(2) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20*, 241−266.

(3) Sahoo, S.; Adhikari, C.; Kuanar, M.; Mishra, B. A Short Review of the Generation of Molecular Descriptors and Their Applications in

Quantitative Structure Property/Activity Relationships. *Curr. Comput.-Aided Drug Des.* **2016**, *12*, 181−205.
(4) Dearden, J. C.; Rotureau, P.; Fayet, G. QSPR prediction of physico-chemical properties for REACH. *SAR QSAR Environ. Res.* **2013**, *24*, 279−318.
(5) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.
(6) Li, Z.; Sillanpää, M. J. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theor. Appl. Genet.* **2012**, *125*, 419−435.
(7) Bishop, C. M. *Pattern recognition and machine learning*; Springer: New York, 2006.
(8) Bruce, C. L.; Melville, J. L.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219−227.
(9) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150−158.
(10) Natekin, A. K. Gradient boosting machines, a tutorial. *Front. Neurorob.* **2013**, *7*, 21.
(11) Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. **2016**, available at arXiv:1603.02754.
(12) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. Y. LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3149−3157.
(13) Meng, Q.; Ke, G.; Wang, T.; Chen, W.; Ye, Q.; Ma, Z. M.; Liu, T. Y. A communication-efficient parallel algorithm for decision tree. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 1279−1287.
(14) Zhang, H.; Si, S.; Hsieh, C. J. GPU acceleration for large-scale tree boosting. *SysML Conference*, 2018.
(15) Katritzky, A.; Fara, D.; Petrukhin, R.; Tatham, D.; Maran, U.; Lomaka, A.; Karelson, M. The Present Utility and Future Potential for Medicinal Chemistry of QSAR / QSPR with Whole Molecule Descriptors. *Curr. Top. Med. Chem.* **2002**, *2*, 1333−1356.
(16) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.
(17) Yada, A.; Nagata, K.; Ando, Y.; Matsumura, T.; Ichinoseki, S.; Sato, K. Machine Learning Approach for Prediction of Reaction Yield with Simulated Catalyst Parameters. *Chem. Lett.* **2018**, *47*, 284−287.
(18) 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*, 186−190.
(19) Dorel, R.; Grugel, C. P.; Haydl, A. M. The Buchwald-Hartwig Amination After 25 Years. *Angew. Chem., Int. Ed.* **2019**, *58*, 17118−17129.
(20) Miyaura, N.; Suzuki, A. Palladium-Catalyzed Cross-Coupling Reactions of Organoboron Compounds. *Chem. Rev.* **1995**, *95*, 2457−2483.
(21) Dhital, R. N.; Sen, A.; Sato, T.; Hu, H.; Ishii, R.; Hashizume, D.; Takaya, H.; Uozumi, Y.; Yamada, Y. M. A. Activator-Promoted Aryl Halide-Dependent Chemoselective Buchwald-Hartwig and Suzuki-Miyaura Type Cross-Coupling Reactions. *Org. Lett.* **2020**, *22*, 4797−4801.
(22) RDKit. Open-source cheminformatics. http://rdkit.org/ (accessed 3 September, 2021).
(23) List of Available Descriptors. https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors (accessed 3 September, 2021).
(24) *Pymatgen (Python Materials Genomics)*. https://pymatgen.org/ (accessed 3 September, 2021).
(25) *Spartan'20*; Wavefunction, Inc., https://www.wavefun.com/ (accessed 3 September, 2021).
(26) Kaneko, H. Data Visualization, Regression, Applicability Domains and Inverse Analysis Based on Generative Topographic Mapping. *Mol. Inf.* **2019**, *38*, 1800088.
(27) Kanno, Y.; Kaneko, H. Improvement of Predictive Accuracy in Semi-Supervised Regression Analysis by Selecting Unlabeled Chemical Structures. *Chemom. Intell. Lab. Syst.* **2019**, *191*, 82−87.
(28) https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule (accessed 3 September, 2021).
(29) Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated double cross validation. *J. Chemom.* **2009**, *23*, 160−171.
(30) Zhou, T.; Xie, P.-P.; Ji, C.-L.; Hong, X.; Szostak, M. Decarbonylative Suzuki-Miyaura Cross-Coupling of Aroyl Chlorides. *Org. Lett.* **2020**, *22*, 6434−6440.
(31) Bhunia, S.; Kumar, S. V.; Ma, D. N,N′-Bisoxalamides Enhance the Catalytic Activity in Cu-Catalyzed Coupling of (Hetero)Aryl Bromides with Anilines and Secondary Amines. *J. Org. Chem.* **2017**, *82*, 12603−12612.
(32) Janke, J.; Ehlers, P.; Villinger, A.; Langer, P. Regioselective Synthesis of Thieno[3,2-b]quinolones by Acylation/Two-Fold Buchwald-Hartwig Reactions. *Eur. J. Org. Chem.* **2019**, 7255−7263.
(33) Aoki, Y.; Toyoda, T.; Kawasaki, H.; Takaya, H.; Sharama, A. K.; Morokuma, K.; Nakamura, M. Iron-Catalyzed Chemoselective C−N Coupling Reaction: A Protecting-Group-Free Amination of Aryl Halides Bearing Amino or Hydroxy Groups. *Asian J. Org. Chem.* **2020**, *9*, 372.
(34) ChemCupid. online compound structure search. Namiki Shoji, Inc. https://www.namiki-s.co.jp/compound/ (accessed 3 September, 2021).
(35) Marvin. https://chemaxon.com/products/marvin (accessed 3 September, 2021).

https://doi.org/10.1021/acsomega.1c04826
ACS Omega 2021, 6, 27578−27586