25 August 2025

Construction of an Interpretable Machine Learning Model for Yield Prediction and Mechanistic Elucidation Enabled by Global Reaction Route Mapping

Takahiro Doba[1], Yu Harabuchi[2], Yuuya Nagata[2], Satoshi Maeda[2]

1. Kyoto University
2. Hokkaido University

Abstract

In the past decade, machine learning has emerged as a powerful tool to predict reaction outcomes. However, mechanistic interpretability of the constructed machine learning models remains limited due to the use of domain-specific and often arbitrary descriptors. Herein we demonstrate that an energy descriptor comprising the energies of the possible intermediates in the reaction system serves as a suitable descriptor for the construction of an interpretable machine learning model for mechanistic elucidation. The energy descriptor was calculated using the single-component artificial force induced reaction (SC-AFIR) method, which autonomously and comprehensively searches for intermediates of a target reaction, and subsequently used to train machine learning models for reaction yield prediction. Linear models with regularization showed good predictions for the hold-out samples (RMSE < 7% yield) and the coefficients of the models provided information on how the energies of the intermediates should be adjusted to enhance the reaction yield. This work highlights the utility of energy descriptors in constructing mechanistically interpretable machine learning models for predictive tasks in chemistry.

Keywords

Artificial force induced reaction method, Reaction path network, Energy descriptor, Robot assisted experiment, Linear models, Yield prediction, Mechanistic elucidation

Posted on 25 August 2025 — CC-BY-NC 4.0 — This is a preprint and has not been peer reviewed. Data may be preliminary. — https://doi.org/10.26434/chemrxiv-2025-0qr26

Title

Construction of an Interpretable Machine Learning Model for Yield Prediction and Mechanistic Elucidation Enabled by Global Reaction Route Mapping

Authors

Takahiro Doba,*[1] Yu Harabuchi,[2,3] Yuuya Nagata,[2,3] Satoshi Maeda*[2,3,4]

[1]*International Research Center for Elements Science, Institute for Chemical Research, Kyoto University*

[2]*Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University*

[3]*ERATO Maeda Artificial Intelligence for Chemical Reaction Design and Discovery Project, Hokkaido University*

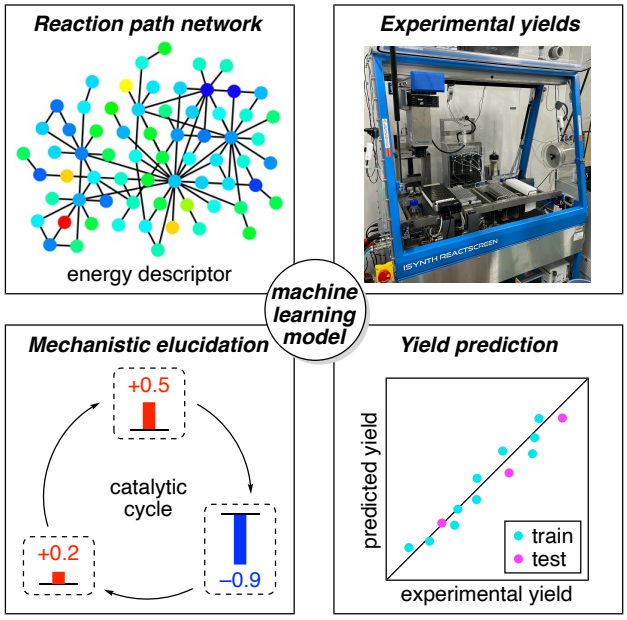
[4]*Department of Chemistry, Faculty of Science, Hokkaido University*

Abstract

In the past decade, machine learning has emerged as a powerful tool to predict reaction outcomes. However, mechanistic interpretability of the constructed machine learning models remains limited due to the use of domain-specific and often arbitrary descriptors. Herein we demonstrate that an energy descriptor comprising the energies of the possible intermediates in the reaction system serves as a suitable descriptor for the construction of an interpretable machine learning model for mechanistic elucidation. The energy descriptor was calculated using the single-component artificial force induced reaction (SC-AFIR) method, which autonomously and comprehensively searches for intermediates of a target reaction, and subsequently used to train machine learning models for reaction yield prediction. Linear models with regularization showed good predictions for the hold-out samples (RMSE < 7% yield) and the coefficients of the models provided information on how the energies of the intermediates should be adjusted to enhance the reaction yield. This work highlights the utility of energy descriptors in constructing mechanistically interpretable machine learning models for predictive tasks in chemistry.

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

TOC



Main Text

Introduction

Over the past decade, machine learning has emerged as a powerful tool for predicting reaction outcomes and accelerating reaction optimization.[1] In a typical machine learning workflow, input molecules are first transformed into feature vectors.[2] Molecular fingerprints, along with physical, chemical, and quantum chemical features, are commonly employed for this vectorization.[3] More advanced approaches incorporate molecular stereostructure and graph-based message passing techniques.[4] Additionally, SMILES representations have played a key role in enabling the use of language models for reaction outcome prediction.[5] These molecular descriptors are employed as inputs for machine learning models, which are trained by minimizing predefined loss functions to accurately predict target variables. A range of models including linear regressors, decision trees, and neural networks are frequently used to predict reaction selectivity and yield. However, the mechanistic interpretability of these models remains limited, owing to the use of domain-specific and often heuristically chosen descriptors. Although the feature importance can be extracted from the machine learning model for interpretation, these values do not necessarily correspond directly to chemically meaningful aspects of the reaction.[6]

Based on this underlying problems associated with interpretability, we sought to

https://doi.org/10.26434/chemrxiv-2025-0qr26 ORCID: https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

investigate the potential of using the energies of the possible intermediates in the reaction system as descriptor. This design is based on the empirical Bell (Brønsted)-Evans-Polanyi (BEP) relationship, which assumes a linear relationship between the activation energy and the free energy, in which kinetic information is stipulated based on thermodynamic information.[7] This approach would simplify the workflow by circumventing the use of transition state energies.[8] Despite this straightforward design, the utilization of descriptors constructed by the energies of the intermediates has been a great challenge due to the difficulty of determining the energies of the possible intermediates present in the reaction mixture. To obtain the "energy descriptor", we decided to use the single component-artificial force induced reaction (SC-AFIR) method, available through the Global Reaction Route Mapping (GRRM) program,[9] which is one of the most powerful methods to search reaction paths.[10] Starting from a single input file specifying the initial reactants, the program iteratively selects two fragments within an intermediate and applies an artificial force to induce a reaction and generate a reaction path. This iterative process, guided by kinetic navigation, results in the construction of a network of realistic reaction paths. The resulting reaction path network contains a sufficient number of relevant intermediates to describe the outcome of the target reaction. By using this energy descriptor for machine learning, the feature importance derived from the machine learning model can be directly associated with specific intermediates, thereby providing insights into which species play critical roles in the reaction mechanism.

Figure 1 outlines a workflow of the construction of an interpretable machine learning model for yield prediction utilizing the energy descriptor calculated by the SC-AFIR method. First, the SC-AFIR method was applied to the initial reactants of a model reaction to obtain the reaction path network of energetically accessible intermediates. Next, substituents were introduced to the identified intermediates, followed by single-point energy calculations and geometry optimizations. The resulting energies were organized into a matrix, with columns representing the intermediates and rows corresponding to the various combinations of substituents. In parallel, an automated robotic platform (Chemspeed SWING) was employed to carry out the reactions with the substituents to afford experimental yields. The obtained yields were aligned to a vector, with its length matching the number of substituent combinations. Then, a machine learning model was constructed between the energy descriptor and the

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

experimental yields. In this work, we preferably investigated linear models such as ordinary least squares (OLS), partial least squares (PLS), Ridge, and Lasso regressions because of the limited number of data points.[11] Leave-one-group-out cross-validation (LOGOCV) was used to assess the ability of the model to predict reaction yield for the holdout samples.[12] Furthermore, the number of explanatory variables was reduced by recursive feature elimination (RFE) so that less important intermediates for yield prediction would be excluded in the subsequent steps.[13] Finally, the intermediates that play key roles in determining the reaction yield were identified based on the coefficients of the linear model. This workflow enables the construction of a machine learning model capable of predicting reaction yields while simultaneously providing mechanistic insight through the analysis of feature importance.

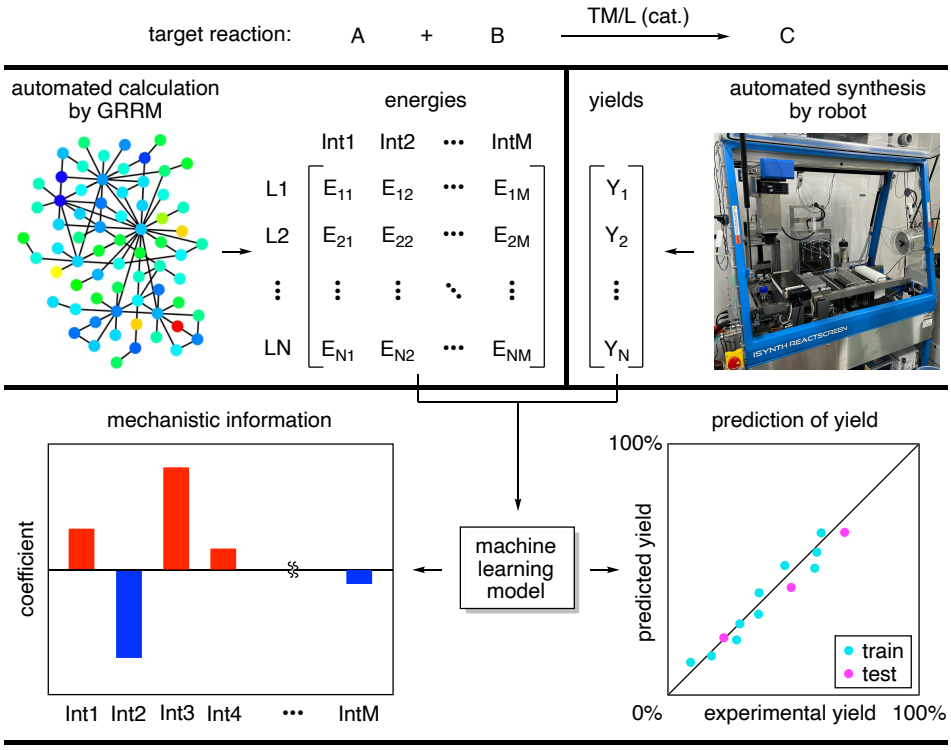


Figure 1. The workflow of using the energy descriptor for the construction of machine learning models and the extraction of mechanistic information. The automated calculation by GRRM identifies intermediates (Int1, Int2, … IntM) and their energies with different substituents (E_{n1}, E_{n2}, … E_{nM}). A robotic system performs the reaction, providing experimental yields (Y_1, Y_2, … Y_N).

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Results and Discussion

To demonstrate the effectiveness of this workflow, we first selected a model reaction that is easy to perform and analyze. We applied this approach to the palladium-catalyzed oxidative C–H alkenylation of simple arenes with electron-deficient alkenes using amino acid and pyridone as ligands, both of which are commercially available.[14] The dual-ligand system facilitated rapid screening of multiple reactions through the preparation of stock solutions for each ligand. 1 equiv of *o*-xylene was reacted with 2 equiv of ethyl acrylate in the presence of 10 mol % of $Pd(OAc)_2$ as a catalyst, 20 mol % of mono-protected amino acid (MPAA) and 20 mol % of pyridone as ligands, and 3 equiv of AgOAc as an oxidant in HFIP at 100 °C for 24 h under air in the well of the robotic platform. Electronically and sterically diverse 8 MPAAs and 13 pyridones were selected based on their commercial availability. The oxidative C–H alkenylation product was obtained in low to high yields with excellent *E*-selectivity and over 94% β-selectivity.

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

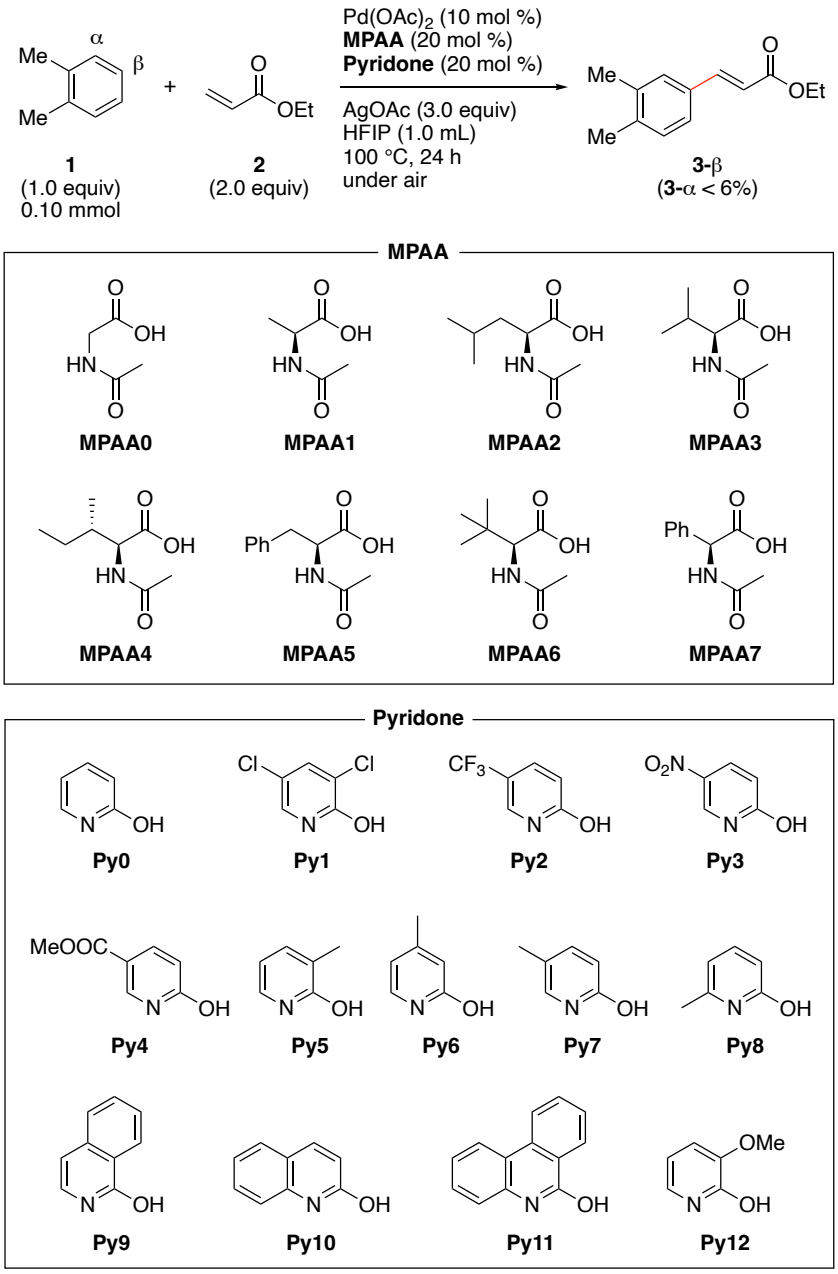


Figure 2. Palladium-catalyzed oxidative C–H alkenylation of arenes with electron-deficient alkenes using amino acid and pyridone as ligands. 8 MPAAs and 13 pyridones were selected by commercial availability.

Next, we applied the SC-AFIR method to the palladium-catalyzed reaction. The input file for the model reaction consisted of a molecule of benzene, methyl acrylate, palladium/MPAA complex, and pyridone as the initial reactants. The calculation was conducted using ONIOM at the level of B3LYP/LanL2DZ for the high layer and UFF for the low layer.[15,16] 14 atoms at which the reaction may take place such as palladium

https://doi.org/10.26434/chemrxiv-2025-0qr26 ORCID: https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

and one of the C–H bonds (as indicated by the red shading in Figure 3a) were designated as the high layer and others were designated as the low layer. All reaction path searches were guided by the kinetics-based navigation method, where the equilibrium structures (EQs) to which the next SC-AFIR procedure is applied were selected based on the "traffic volume" determined by the rate constant matrix contraction (RCMC) method (see Supporting Information for details).[17] A weak universal force was applied to each intermediate to maintain reasonable intermolecular proximity. To obtain the energetically accessible intermediates, the SC-AFIR method was applied by the following multi-step procedure, in which the kinetics-based navigation was altered several times by successively assigning the population to the EQs of putative reaction intermediates. First, 500 effective AFIR paths were computed from 111 sets of randomly generated initial structures. In the kinetic analyses, the initial population ($1/n$) was assigned to n initial structures. Then, the population of 1.0 was assigned to the C–H activation complex with the lowest electronic energy in the reaction path network, and 500 effective AFIR paths were further computed to obtain the alkene coordination step. Similarly, the population was assigned to the alkene coordination complex with the lowest electronic energy, and 500 effective AFIR paths were computed to obtain the alkene insertion step. The same procedure was applied to the proton transfer, carbopalladation, and β-H elimination steps. For simplicity, we did not explicitly include HFIP and Ag(OAc) molecules for the calculation. In total, 2500 effective AFIR paths were computed to obtain the reaction path network containing the productive reaction paths giving the cross-coupling product and other unproductive paths. The reaction path network constructed using this multi-step method contained 3954 intermediates and 7120 transition states including structural conformers (Figure 3a). By employing the traffic volume and the minimum path algorithms,[17] 52 key intermediates that are kinetically and thermodynamically accessible under the reaction conditions were identified.

In calculating the energy descriptor, the intermediates were decomposed into individual components according to the connectivity of the atoms, which was determined using a distance matrix and a predefined threshold based on atomic radii. The separated components were subjected to single-point energy calculation and geometry optimization, followed by the introduction of a set of substituents. During this process, multiple conformers were generated; for example, the introduction of an

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

isopropyl group resulted in three conformers corresponding to 120° rotations around the C–C bond. Subsequently, the substituted components were subjected to single-point energy calculations and geometry optimizations to determine their respective energies (E_a, E_b, E_c in Figure 3b). Finally, these energies were summed to obtain a single energy value (i.e. $E = E_a + E_b + E_c$) and used to construct the energy descriptor. This procedure was applied automatically to the 52 key intermediates and all combinations of substituents using a Python program.

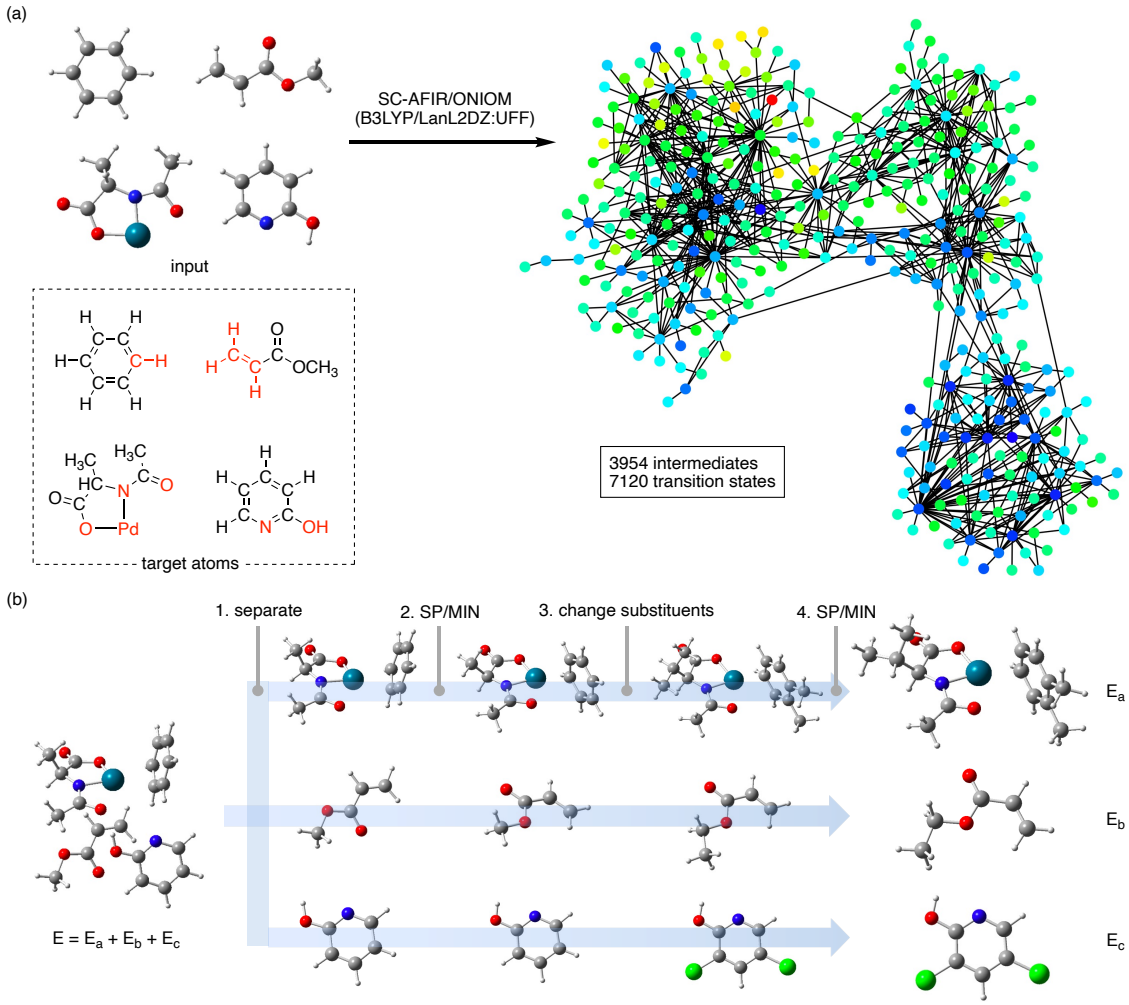


Figure 3. Theoretical calculation and calculation of the energy descriptor. (a) Automated reaction path network exploration using SC-AFIR in the GRRM program. Target atoms shown in red indicate the atoms to apply the artificial forces. Nodes and edges of the graph represent intermediates and reaction paths, respectively. (b) Stepwise calculation of the energy used for the construction of the energy descriptor. SP: single-point energy calculation, MIN: geometry optimization.

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

The reactions for all combinations of MPAAs and pyridones were carried out using the Chemspeed SWING robotic platform. The heatmap in Figure 4a shows the average yields of two runs of the oxidative C–H alkenylation product **3**-β. Variations in MPAAs and pyridones are represented along the rows and columns, respectively, with observed yields ranging from 17% to 70%. Reactions were carried out in wells containing small vials, and GC samples were prepared autonomously by passing the reaction mixtures through a pad of silica gel (Figure 4b).

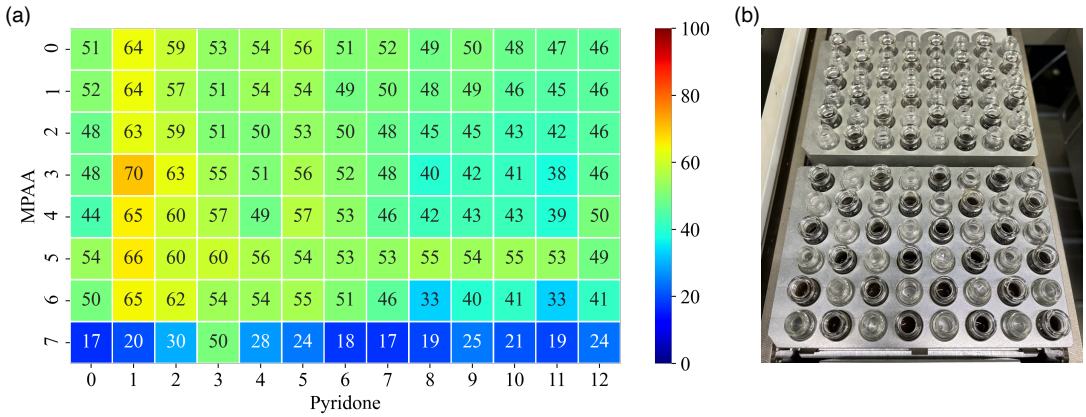


Figure 4. The model reactions performed on a robotic platform. (a) Experimental yields of the oxidative C–H alkenylation product **3**-β. The heatmap shows the average yields of two runs using various combinations of MPAA (rows) and pyridone (columns). (b) The reaction wells of Chemspeed SWING robotic platform used for automating the synthesis and screening process.

With the energy descriptors and corresponding experimental yields in hand, we constructed machine learning models to predict reaction yields for holdout samples and to gain insights into the mechanism of the reaction. All energy values included in the energy descriptor were referenced to the energies of the unreacted intermediates. Due to the limited number of data points, we focused on linear models—ordinary least squares (OLS), partial least squares (PLS), Ridge, and Lasso regression. Each column of the energy descriptor was standardized to unit variance so that the coefficient values can be compared directly. These simplified models also contribute to maintaining the interpretability of the model. Leave-one-group-out cross-validation (LOGOCV) was applied based on pyridone subgroups: one pyridone group was held out for testing, while the remaining groups were used for training, and the holdout group was used to

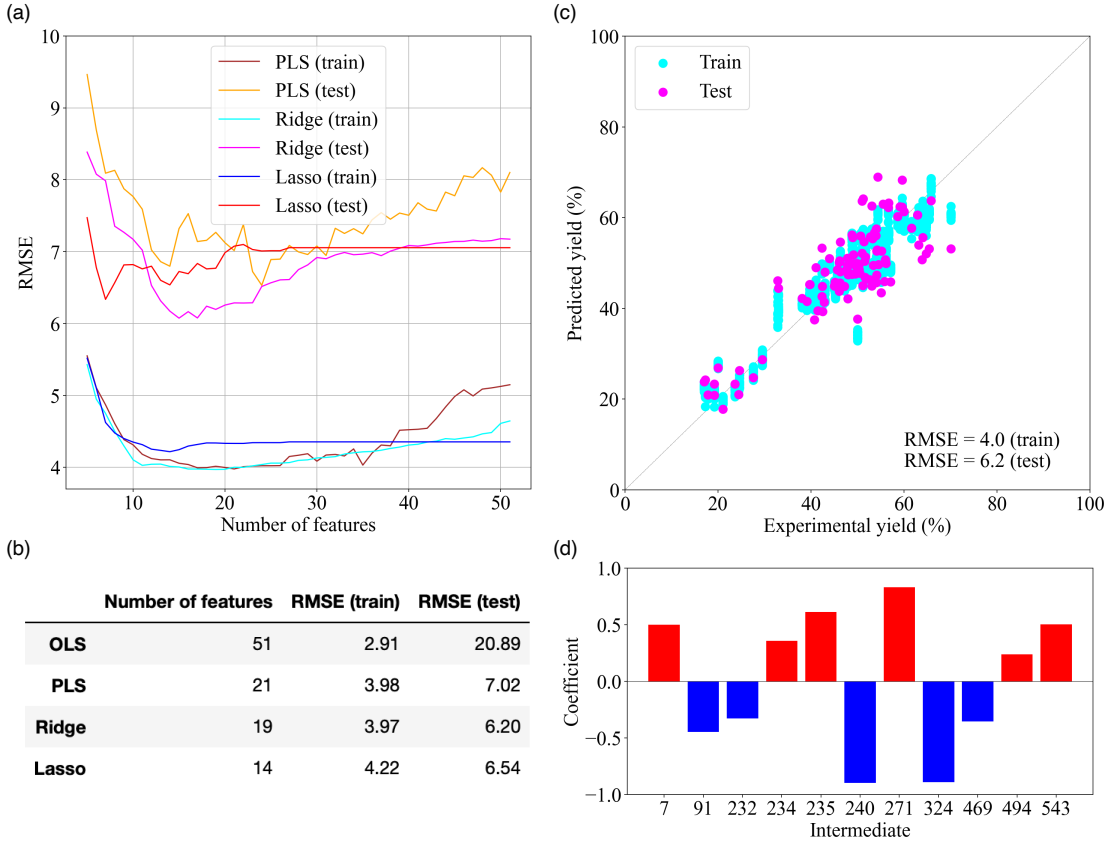
https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

assess model performance. To further improve model performance and interpretability, recursive feature elimination (RFE) was employed to reduce the number of explanatory variables. In RFE, a linear model was first trained using all features (i.e., all columns of the energy descriptor). The feature with the smallest absolute coefficient was removed, and a new model was trained on the reduced set. This process was repeated iteratively, allowing the model to achieve high predictive accuracy with a minimal set of features. Figure 5a shows the root mean squared error (RMSE) for both training and test sets across the linear models. The optimal number of features was identified at the point where the RMSE for the training set was minimized. The RMSE values for the test sets, calculated at the optimal feature number, are summarized in Figure 5b. Among the models tested, the OLS model exhibited the lowest RMSE for the training set when using all features; however, its RMSE for the test set was substantially higher, reflecting poor generalization performance due to the absence of regularization. In contrast, the PLS model reduced the RMSE for the test set to 7.02% using 21 features, or energies of 21 intermediates. The Ridge model, incorporating L2 regularization, further lowered the RMSE for the test set to 6.20% with 19 features, indicating that this model can predict the reaction yield of an unknown pyridone with an accuracy of approximately ±6% yield. The Lasso model, incorporating L1 regularization, reduced the number of selected features to 14, although the RMSE for the test set did not improve further. Figure 5c shows a plot between the experimental yields and the yields predicted by the Ridge regression model. The data points exhibit close alignment along the diagonal line, indicating strong agreement between the predicted and experimental yields across the entire range, from low to high yields.

Finally, a Ridge model was trained using the full dataset, from which 11 features were selected using RFE. The corresponding model coefficients are shown in Figure 5d. Energy contributions of intermediates are represented as either positive or negative, offering a clear picture on how the energies of the intermediates should be manipulated. The positive coefficients (red bars) indicate that higher energy values for the corresponding intermediates are associated with higher reaction yields, while negative coefficients (blue bars) suggest that lower energy values are correlated with higher yields. For example, increasing the energy of intermediate 271 correlates with higher yield, while reducing the energy of intermediate 324 improves reaction performance. In both cases, larger absolute values imply that the reaction yields are more strongly

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

influenced by the energy of the corresponding intermediate.



	Number of features	RMSE (train)	RMSE (test)
OLS	51	2.91	20.89
PLS	21	3.98	7.02
Ridge	19	3.97	6.20
Lasso	14	4.22	6.54

Figure 5. Construction of machine learning models. (a) Root mean squared errors (RMSE) of linear models for train and test sets. (b) RMSE for the test sets at the number of features that gave the smallest RMSE for the training sets. (c) Comparison between experimental and predicted yields. (d) Coefficients of the Ridge model trained over all data. X axis indicates the number of intermediates. Python programs were used for the machine learning studies.

The feature importance derived from the machine learning model provided insights into the reaction mechanism. Figure 6 illustrates a plausible catalytic cycle constructed based on the intermediates identified from the coefficients of the Ridge model. The values displayed beneath the intermediate numbers represent the coefficients assigned by the model. The intermediates shown without coefficients were not selected by the model but were included to help connect the selected intermediates. Starting from the initial reactants (**Int-0**), pyridone coordinated to palladium to form a palladium complex bearing two ligands (**Int-7**). The coefficient of **Int-7** was +0.50, suggesting that a higher

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

energy of **Int-7** is associated with a higher reaction yield. This observation aligns with the notion that **Int-7** should not be overly stable in order for the subsequent C–H activation step to proceed efficiently. Although the C–H activation complex (**Int-46**) was not selected by the model, the coefficient of the intermediate after the olefin coordination step (**Int-91**) was –0.45, suggesting that a lower energy of **Int-91** is associated with a higher reaction yield. This observation is consistent with the understanding that efficient olefin coordination is essential for smooth reaction progression. **Int-46** may not have been selected because the energy levels of the intermediates preceding and following the C–H activation step (**Int-7** and **Int-91**, respectively) played a more significant role in determining the reaction yields. The coefficients of **Int-232** and **Int-240** were –0.33 and –0.90, respectively, suggesting that these intermediates need to be stabilized. These intermediates feature dechelated monodentate coordination of the MPAA ligand. This is in line with the fundamental understanding that two substrates must occupy two coordination sites on the square-planar palladium center, positioned *cis* to each other, for the subsequent carbometallation step to proceed. On the other hand, the coefficients of other types of dechelated palladium intermediates (**Int-234** and **Int-235**) were +0.36 and +0.61, respectively, suggesting that these intermediates need to be destabilized. This subtle difference is nontrivial, and the method provides valuable insights into the nuanced dynamics of ligand coordination. The coefficients of the intermediates preceding and following the proton transfer step (**Int-240** and **Int-324**, respectively), in which the imidic acid moiety of the MPAA ligand is converted to an amide, were both strongly negative, suggesting that the proton transfer step needs to be facile. This is consistent with the typical use of a protic HFIP solvent under these reaction conditions.[14] Although **Int-240** and **Int-324** had the two most negative coefficients (–0.90 and –0.89, respectively) and suggest that these intermediates are involved in the turnover-limiting step, these values might include artifacts. This is because the theoretical calculations did not explicitly account for the effects of HFIP, and the energy descriptor may lack sufficient information to fully capture the energetics of the proton transfer step. In contrast, the coefficient of **Int-271** was strongly positive (+0.83), indicating that the ligand dissociation is not preferred. From **Int-324**, carbopalladation occurred to give **Int-357**, in which the phenyl group is oriented toward the palladium center. The coefficient of the intermediate following the C–C bond rotation (**Int-469**) was –0.36,

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

suggesting that a lower energy of **Int-469** is associated with a higher reaction yield. This is in accordance with the requirement that the β-hydrogen must be oriented toward the palladium center for the subsequent β-hydride elimination step to proceed. Interestingly, the coefficient for the product complex (**Int-494**) was positive (+0.24), suggesting that a higher energy of **Int-494** is associated with a higher reaction yield. This is consistent with the idea that the product complex must be sufficiently unstable to facilitate product dissociation from the palladium center and enable catalyst regeneration. Conversely, the coefficient of **Int-543** was positive (+0.50), indicating that the ligand dissociation should not occur.

Without relying on arbitrarily chosen descriptors, the model successfully identified key intermediates within the catalytic cycle and off-cycle species that play significant roles in determining reaction yield. The resulting coefficients were interpretable and consistent with fundamental understanding of palladium-catalyzed C–H activation and Mizoroki-Heck reactions.

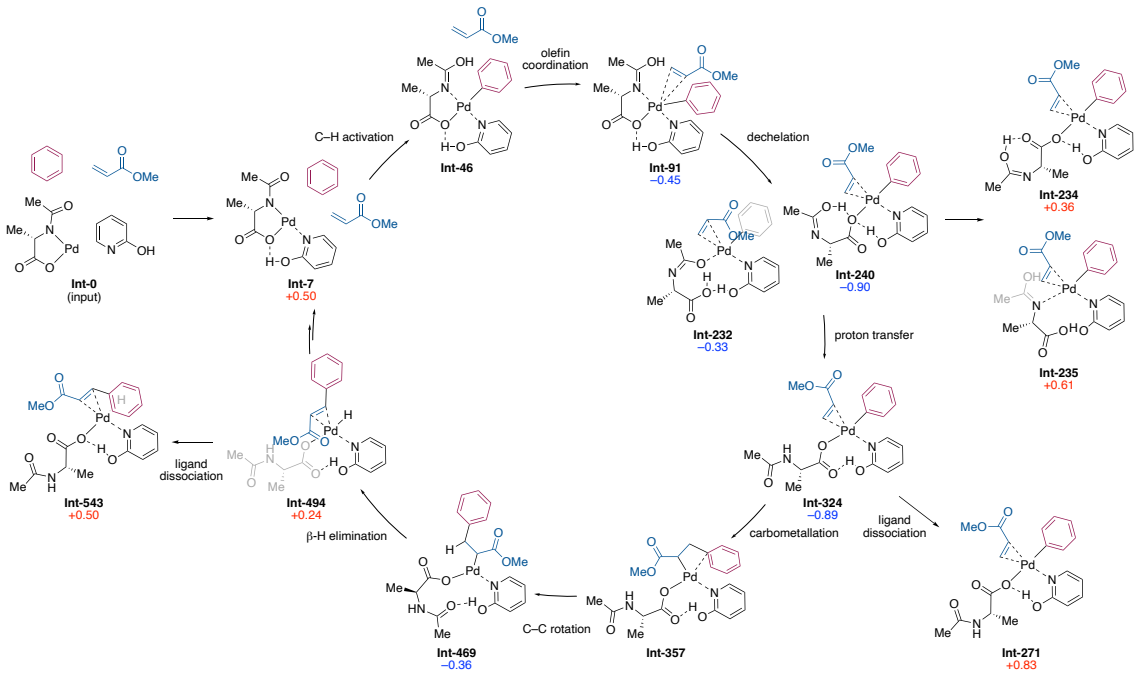


Figure 6. A plausible catalytic cycle constructed based on the intermediates identified from the coefficients of the Ridge model.

Conclusions

In summary, an interpretable machine learning model that can provide insight into

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

the mechanism of the reaction was constructed using the energy descriptor automatically generated by the SC-AFIR method. Linear models performed effectively, with the Ridge model providing accurate predictions for the holdout samples (RMSE < 7% yield). The coefficients of the Ridge model provided information on how the energies of the intermediates should be adjusted to enhance the reaction yield. This work highlights the synergy of this integrative approach combining theoretical calculation, automated synthesis, and machine learning, offering a powerful framework for predicting reaction yields and obtaining mechanistic insights. Extension of this method to identify improved catalysts and reaction conditions is currently underway in our laboratory.

Associated Content

Data Availability Statement

The source codes for the energy descriptor calculations and machine learning are available at https://github.com/takahiro-doba-research/maeda-lab-postdoc

Supporting Information

The Supporting Information is available free of charge on the ChemRxiv website.

Procedures for theoretical calculations, robotic experiments, and machine learning (PDF)

Author Information

Corresponding Authors

Takahiro Doba – *International Research Center for Elements Science, Institute for Chemical Research, Kyoto University, Kyoto, 611-0011, Japan;* orcid.org/0000-0003-3593-9723; Email: doba.takahiro.3t@kyoto-u.ac.jp

Satoshi Maeda – *Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo 060-0810, Japan; Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Sapporo 001-0021, Japan; Japan Science and Technology Agency (JST), Exploratory Research for Advanced Technology (ERATO), Maeda Artificial Intelligence for Chemical Reaction Design and Discovery Project, Sapporo 060-0810, Japan;* orcid.org/0000-0001-8822-1147; Email: smaeda@eis.hokudai.ac.jp

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Authors

Yu Harabuchi − *Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Sapporo 001-0021, Japan; Japan Science and Technology Agency (JST), Exploratory Research for Advanced Technology (ERATO), Maeda Artificial Intelligence for Chemical Reaction Design and Discovery Project, Sapporo 060-0810, Japan;* orcid.org/0000-0001-8313-3236; Email: y_harabuchi@icredd.hokudai.ac.jp

Yuuya Nagata − *Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Sapporo 001-0021, Japan; Japan Science and Technology Agency (JST), Exploratory Research for Advanced Technology (ERATO), Maeda Artificial Intelligence for Chemical Reaction Design and Discovery Project, Sapporo 060-0810, Japan;* orcid.org/0000-0001-5926-5845; Email: nagata@icredd.hokudai.ac.jp

Author Contributions

The manuscript was written with the contributions of all authors.

Notes

The authors declare no competing financial interest.

Acknowledgements

This study was supported by JST-ERATO (grant number JPMJER1903) to S.M., JSPS KAKENHI (grant number 22KJ0043) to T.D., and JSPS KAKENHI (grant number 23B203) to Y.N. and Y.H.

References

[1] (a) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443. (b) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-Catalyzed C–C Bond Formation Enabled through Ligand Parameterization. *Science* **2018**, *362* (6415), 670–674. (c) Granda, J. M.; Donina, L.;

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Dragone, V.; Long, D.-L.; Cronin, L. Controlling an Organic Synthesis Robot with Machine Learning to Search for New Reactivity. *Nature* **2018**, *559* (7714), 377–381. (d) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. (e) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. (f) 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* (6385), 186–190. (g) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363* (6424), eaau5631. (h) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Molga, K.; Mlynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational Planning of the Synthesis of Complex Natural Products. *Nature* **2020**, *588* (7836), 83–88. (i) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379–1390. (j) Funes-Ardoiz, I.; Schoenebeck, F. Established and Emerging Computational Tools to Study Homogeneous Catalysis—From Quantum Mechanics to Machine Learning. *Chem* **2020**, *6* (8), 1904–1913. (k) Hueffel, J. A.; Sperger, T.; Funes-Ardoiz, I.; Ward, J. S.; Rissanen, K.; Schoenebeck, F. Accelerated Dinuclear Palladium Catalyst Identification through Unsupervised Machine Learning. *Science* **2021**, *374* (6571), 1134–1140. (l) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. *Chem. Sci.* **2021**, *12* (3), 1163–1175.

[2] Pattanaik, L.; Coley, C. W. Molecular Representation: Going Long on Fingerprints. *Chem* **2020**, *6* (6), 1204–1207.

[3] Boldini, D.; Ballabio, D.; Consonni, V.; Todeschini, R.; Grisoni, F.; Sieber, S. A. Effectiveness of Molecular Fingerprints for Exploring the Chemical Space of Natural Products. *J. Cheminformatics* **2024**, *16* (1), 35.

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

[4] (a) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148* (24), 241722. (b) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388.

[5] Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn.: Sci. Technol.* **2021**, *2* (1), 015016.

[6] (a) Dybowski, R. Interpretable Machine Learning as a Tool for Scientific Discovery in Chemistry. *New J. Chem.* **2020**, *44* (48), 20914–20920. (b) Oviedo, F.; Ferres, J. L.; Buonassisi, T.; Butler, K. T. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. *Acc. Mater. Res.* **2022**, *3* (6), 597–607.

[7] (a) Brönsted, J. N. Acid and Basic Catalysis. *Chem. Rev.* **1928**, *5* (3), 231–338. (b) Evans, M. G.; Polanyi, M. Some Applications of the Transition State Method to the Calculation of Reaction Velocities, Especially in Solution. *Trans. Faraday Soc.* **1935**, *31*, 875. (c) Evans, M. G.; Polanyi, M. Further Considerations on the Thermodynamics of Chemical Equilibria and Reaction Rates. *Trans. Faraday Soc.* **1936**, *32*, 1333. (d) Bell, R. P. The Theory of Reactions Involving Proton Transfers. *Proc. R. Soc. Lond. A* **1936**, *154* (882), 414–429. (e) Evans, M. G.; Polanyi, M. Inertia and Driving Force of Chemical Reactions. *Trans. Faraday Soc.* **1938**, *34*, 11.

[8] (a) Roy, S.; Goedecker, S.; Hellmann, V. Bell-Evans-Polanyi Principle for Molecular Dynamics Trajectories and Its Implications for Global Optimization. *Phys. Rev. E* **2008**, *77* (5), 056707. (b) Vojvodic, A.; Hellman, A.; Ruberto, C.; Lundqvist, B. I. From Electronic Structure to Catalytic Activity: A Single Descriptor for Adsorption and Reactivity on Transition-Metal Carbides. *Phys. Rev. Lett.* **2009**, *103* (14), 146103. (c) Santos, E.; Lundin, A.; Pötting, K.; Quaino, P.; Schmickler, W. Model for the Electrocatalysis of Hydrogen Evolution. *Phys. Rev. B* **2009**, *79* (23), 235436. (d) Sutton, J. E.; Vlachos, D. G. A Theoretical and Computational Analysis of Linear Free Energy Relations for the Estimation of Activation Energies. *ACS Catal.* **2012**, *2* (8), 1624–1634. (e) Wang, S.; Vorotnikov, V.; Sutton, J. E.; Vlachos, D. G. Brønsted–Evans–Polanyi and Transition State Scaling Relations of Furan Derivatives on Pd(111) and Their Relation to Those of Small Molecules. *ACS Catal.* **2014**, *4* (2), 604–612. (f) Zhdanov, V. P. Late

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Stage of the Formation of a Protein Corona around Nanoparticles in Biofluids. *Phys. Rev. E* **2022**, *105* (1), 014402. (g) Yang, T. T.; Saidi, W. A. The Bell-Evans-Polanyi Relation for Hydrogen Evolution Reaction from First-Principles. *npj Comput. Mater.* **2024**, *10* (1), 98.

[9] (a) Maeda, S.; Taketsugu, T.; Morokuma, K. Exploring Transition State Structures for Intramolecular Pathways by the Artificial Force Induced Reaction Method. *J. Comput. Chem.* **2014**, *35* (2), 166–173. (b) Maeda, S.; Harabuchi, Y.; Takagi, M.; Saita, K.; Suzuki, K.; Ichino, T.; Sumiya, Y.; Sugiyama, K.; Ono, Y. Implementation and Performance of the Artificial Force Induced Reaction Method in the GRRM17 Program. *J. Comput. Chem.* **2018**, *39*, 233–251. (c) Maeda, S.; Harabuchi, Y. Exploring Paths of Chemical Transformations in Molecular and Periodic Systems: An Approach Utilizing Force. *WIREs Comput. Mol. Sci.* **2021**, *11*, e1538.

[10] (a) Schlegel, H. B. Exploring Potential Energy Surfaces for Chemical Reactions: An Overview of Some Practical Methods. *J. Comput. Chem.* **2003**, *24* (12), 1514–1527. (b) Maeda, S.; Ohno, K.; Morokuma, K. Systematic Exploration of the Mechanism of Chemical Reactions: The Global Reaction Route Mapping (GRRM) Strategy Using the ADDF and AFIR Methods. *Phys. Chem. Chem. Phys.* **2013**, *15* (11), 3683. (c) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for Exploring Reaction Space in Molecular Systems. *WIREs Comput. Mol. Sci.* **2018**, *8* (2), e1354. (d) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of Reaction Pathways and Chemical Transformation Networks. *J. Phys. Chem. A* **2019**, *123* (2), 385–399. (e) Margraf, J. T.; Jung, H.; Scheurer, C.; Reuter, K. Exploring Catalytic Reaction Networks with Machine Learning. *Nat. Catal.* **2023**, *6* (2), 112–121. (f) Blum, V.; Asahi, R.; Autschbach, J.; Bannwarth, C.; Bihlmayer, G.; Blügel, S.; Burns, L. A.; Crawford, T. D.; Dawson, W.; de Jong, W. A.; Draxl, C.; Filippi, C.; Genovese, L.; Giannozzi, P.; Govind, N.; Hammes-Schiffer, S.; Hammond, J. R.; Hourahine, B.; Jain, A.; Kanai, Y.; Kent, P. R. C.; Larsen, A. H.; Lehtola, S.; Li, X.; Lindh, R.; Maeda, S.; Makri, N.; Moussa, J.; Nakajima, T.; Nash, J. A.; Oliveira, M. J. T.; Patel, P. D.; Pizzi, G.; Pourtois, G.; Pritchard, B. P.; Rabani, E.; Reiher, M.; Reining, L.; Ren, X.; Rossi, M.; Schlegel, H. B.; Seriani, N.; Slipchenko, L. V.; Thom, A.; Valeev, E. F.; Van Troeye, B.; Visscher, L.; Vlček, V.; Werner, H.-J.; Williams-Young, D. B.; Windus, T. L. Roadmap on Methods and Software for Electronic Structure Based Simulations in Chemistry and Materials.

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Electron. Struct. **2024**, *6* (4), 042501.

[11] (a) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12* (1), 55–67. (b) Wold, H. Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. *J. Appl. Probab.* **1975**, *12* (S1), 117–142. (c) Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Statist. Soc. B* **1996**, *58* (1), 267–288.

[12] Varma, S.; Simon, R. Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinformatics* **2006**, *7* (1), 91.

[13] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **2002**, *46*, 389–422.

[14] (a) Kubota, A.; Emmert, M. H.; Sanford, M. S. Pyridine Ligands as Promoters in $Pd^{II/0}$-Catalyzed C–H Olefination Reactions. *Org. Lett.* **2012**, *14* (7), 1760–1763. (b) Wang, P.; Verma, P.; Xia, G.; Shi, J.; Qiao, J. X.; Tao, S.; Cheng, P. T. W.; Poss, M. A.; Farmer, M. E.; Yeung, K.-S.; Yu, J.-Q. Ligand-Accelerated Non-Directed C–H Functionalization of Arenes. *Nature* **2017**, *551* (7681), 489–493. (c) Chen, H.; Wedi, P.; Meyer, T.; Tavakoli, G.; van Gemmeren, M. Dual Ligand-Enabled Nondirected C−H Olefination of Arenes. *Angew. Chem. Int. Ed.* **2018**, *57* (9), 2497–2501. (d) Zhao, D.; Xu, P.; Ritter, T. Palladium-Catalyzed Late-Stage Direct Arene Cyanation. *Chem* **2019**, *5* (1), 97–107. (e) Sinha, S. K.; Panja, S.; Grover, J.; Hazra, P. S.; Pandit, S.; Bairagi, Y.; Zhang, X.; Maiti, D. Dual Ligand Enabled Nondirected C–H Chalcogenation of Arenes and Heteroarenes. *J. Am. Chem. Soc.* **2022**, *144* (27), 12032–12042.

[15] Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. A New ONIOM Implementation in Gaussian98. Part I. The Calculation of Energies, Gradients, Vibrational Frequencies and Electric Field Derivatives. *J. Mol. Struct. (Theochem),* **1999**, *462*, 1–21.

[16] (a) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035. (b) Casewit, C. J.; Colwell, K. S.; Rappe, A. K. Application of a Universal Force Field to Organic Molecules. *J. Am. Chem. Soc.* **1992**, *114* (25), 10035–10046. (c) Rappe, A. K.; Colwell, K. S.; Casewit, C. J. Application of a Universal Force Field to Metal Complexes. *Inorg.*

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Chem. **1993**, *32* (16), 3438–3450.

[17] (a) Sumiya, Y.; Nagahata, Y.; Komatsuzaki, T.; Taketsugu, T.; Maeda, S. Kinetic Analysis for the Multistep Profiles of Organic Reactions: Significance of the Conformational Entropy on the Rate Constants of the Claisen Rearrangement. *J. Phys. Chem. A* **2015**, *119* (48), 11641–11649. (b) Sumiya, Y.; Taketsugu, T.; Maeda, S. Full Rate Constant Matrix Contraction Method for Obtaining Branching Ratio of Unimolecular Decomposition. *J. Comput. Chem*. **2017**, *38* (2), 101–109. (c) Sumiya, Y.; Maeda, S. A Reaction Path Network for Wöhler's Urea Synthesis. *Chem. Lett.* **2019**, *48*, 47–50. (d) Sumiya, Y.; Maeda, S. Rate Constant Matrix Contraction Method for Systematic Analysis of Reaction Path Networks. *Chem. Lett.* **2020**, *49*, 553–564.

https://doi.org/10.26434/chemrxiv-2025-0qr26 **ORCID:** https://orcid.org/0000-0003-3593-9723 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0