Prediction Method for Reaction Yield of Deuteration of Polyfluoroperylene using Generative AI Techniques

Kazuhiro Takeda,a\* Naoya Ohtsuka,b,c Toshiyasu Suzuki,b and Norie Momiyamab,c

aDepartment of Applied Chemistry and Biochemical Engineering, Shizuoka University, Hamamatsu, Shizuoka, 4328561, Japan

bInstitute for Molecular Science, Okazaki, Aichi, 4448787, Japan

cMolecular Science Program, Graduate Institute for Advanced Studies, SOKENDAI, Okazaki, Aichi, 4448787, Japan

takeda.kazuhiro@shizuoka.ac.jp

Abstract

Deuterated organic electroluminescent materials are gaining interest due to their enhanced luminous efficiency and durability with applications spanning academia and industry (Saito et al., 1994). Perylene is a typical organic molecule for organic light-emitting devices. Deuterated polyfluoroperylene (PFDPR), in which the hydrogen of the polyfluoroperylene is replaced by a deuterium, has potential as a new luminescent material. However, synthesizing PFDPR is challenging due to the complexity and scale of the required deuteration processes. On the other hand, in machine learning, large amounts of data are required to improve the estimation accuracy. Takeda et al. (2023) has proposed the virtual variables-enabled generation of datasets for the prediction of the yield of the iodination reactions of the polyfluoronaphthalenes. Using this method, this study proposes a model to estimate the non-experimental yield of PFDPR with a high accuracy from a small amount of data. The experimental conditions investigated in this study were two variables; i.e., temperature and time, across 16 conditions. While comprehensive data for the polyfluoronaphthalenes were fully available, the polyfluoroperylene data were limited to only 8 conditions. the experimental data from polyfluoronaphthalenes determined the yield prediction of the polyfluoroperylene under untested conditions. This process involved optimization using virtual variables to maximize the coefficient of determination between the actual and predicted yields of PFDPR. The model's efficacy is highlighted by the close alignment of the predicted and actual yields, offering a promising tool for accelerating the PFDPR synthesis research.

**Keywords**: generative artificial intelligence; small data; prediction of reaction conditions; in-silico data generation; digitalization of organic molecules.

* 1. Introduction

Deuterated organic electroluminescent materials have been reported to improve the luminous efficiency and durability, thus are attracting attention in industry (Saito et al., 1994). Perylene is a typical organic molecule for organic light-emitting devices. We have recently developed an efficient synthesis of polyfluoronaphthalene (Ohtsuka et al., 2023, Takeda et al., 2023). Furthermore, the synthetic method has been successfully applied and developed to synthesize polyfluoroperylene (F8), a fluorine-modified in eight positions (in preparation for submission). Deuterated F8 (deuterated polyfluoroperylene; PFDPR), in which the hydrogen of F8 is replaced by a deuterium, has the potential as a new luminescent material. However, the synthesis of F8 requires multiple steps and it is difficult to synthesize large quantities of F8 as it requires numerous deuteration experiments. When assessing the luminescence properties, it is necessary to fully scrutinize whether all four of the hydrogen atoms are deuterated or partially deuterated in the organic material.

On the other hand, machine learning has been used to try to determine the conditions for a chemical reaction (Struble et al., 2020). Many algorithms for machine learning have been proposed. PyCaret (2020) can treat 26 algorithms, such as ‘Linear Regression’, ‘Lasso Regression’, ‘Ridge Regression’, ‘Elastic Net’, ‘Least Angle Regression’, ‘Lasso Least Angle Regression’, ‘Orthogonal Matching Pursuit’, ‘Bayesian Ridge’, ‘Automatic Relevance Determination’, ‘Passive Aggressive Regressor’, ‘Random Sample Consensus’, ‘TheilSen Regressor’, ‘Huber Regressor’, ‘Kernel Ridge’, ‘Support Vector Regression’, ‘K Neighbors Regressor’, ‘Decision Tree Regressor’, ‘Random Forest Regressor’, ‘Extra Trees Regressor’, ‘AdaBoost Regressor’, ‘Gradient Boosting Regressor’, ‘MLP Regressor’, ‘Extreme Gradient Boosting’, ‘Light Gradient Boosting Machine’, ‘CatBoost Regressor’, and ‘Dummy Regressor’. When the experimental data for F8 is limited, the algorithm employing Bayesian optimization can effectively identify the optimal conditions using limited data (Kondo et al., 2023). However, this algorithm does not offer explainable insights into the reaction mechanisms as it primarily processes data through a statistical analysis. Takeda et al. (2023) has proposed a generative machine learning method with virtual variables (GMLV) to predict results under non-experimental conditions (Figure 1). The GMLV can also contribute to the understanding of the reaction mechanism as the virtual variables represent the relationships between the reactants. In this study, we report the development of a predictive model for the deuteration of F8 using the GMLV.

* 1. Generative Machine Learning Method with Virtual Variables

The algorithm of GMLV (Takeda et al., 2023) is illustrated in Figure 1. This algorithm is based on a dual-loop structure. To express the relationship between the objective reactants, virtual variables were assumed in the outer loop. The optimal value of each virtual variable was investigated in the inner loop by maximizing the evaluation value. The Bayesian optimization method, which can determine the optimal value even with a small number of trials (Kondo et al., 2023), was used for the investigations. Therefore, in the outer loop, the explanatory variables were virtual variables, and the objective function was the optimizing inner loop. In the inner loop, the non-experimental yields were assumed and the coefficient of determination R2 for the experimental and assumed non-experimental yields was maximized. As there were multiple assumed yields in the investigation, a dual-annealing optimization method, which is less likely to fall into local solutions (Xiang et al. 1997), was used, and the minimum value obtained after three trials was adopted. The yield surface was predicted using SVR (Vapnik, 1995), which is most suitable for predicting experimental yields. This study developed a more precise regression model by incorporating additional appropriate samples. A reaction space model was constructed, representing yield as a function of time and temperature. This involved using assumed yield data for conditions not experimentally tested. The model's explanatory variables included both the experimented time and temperature, and virtual variables representing the relationship between the reactants. The inner loop's objective function aimed to maximize the R-squared (R²) value, enhancing the predictive accuracy for both the experimental and assumed non-experimental yields. Consequently, the non-experimental yield that optimizes R² is considered most fitting for this model. Additionally, the optimized virtual variables establish a more accurate relationship between the objective reactants.



Figure 1: Flow diagram of the GMLV.

* 1. Objective Reactants and Conditions

Figure 2 illustrates the various fluorinated naphthalenes and perylenes used in the study. 1,2,3,4,5,6,8-Heptafluoronaphthalene, a naphthalene derivative with seven fluorine atoms, is denoted as ‘F7’. There are two types of hexafluoronaphthalenes, modified with six fluorine atoms; i.e., 1,2,4,5,7,8-hexafluoronaphthalene (‘3,6H-F6’) and 1,2,4,5,6,8-hexafluoronaphthalene (‘3,7H-F6’). Similarly, 1,3,4,6,7,9,10,12-octafluoroperylene, a perylene variant with eight fluorine atoms, was labeled as ‘F8’.

In the magnesiation process, polyfluoronaphthalene and polyfluoroperylene were treated with Mg(TMP)2・2LiBr in 0.13 M THF solvent and subsequently deuterated using 0.5 mL D2O. The Mg(TMP)2・2LiBr equivalents used were 1.2 for F7, 2.4 for F6, and 4.8 for F8. In our experiments, both hydrogens in F6 were fully deuterated. For F8, depending on whether one, two, three, or all four hydrogens are deuterated, the compounds were referred to as D1, D2, D3, and D4, respectively. The reactions were conducted at temperatures of -78 °C, -40 °C, 0 °C, and 25 °C, and for durations of 0.5 h, 1.0 h, 2.0 h, or 4.0 h.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| F7 | 3,6H-F6 | 3,7H-F6 | F8 |

Figure 2 Structures of polyfluoronaphthalenes and polyfluoroperylene.

* 1. Results and Discussion

For F6 (both 3,6H-F6 and 3,7H-F6) and F7, the deuteration yields were determined across 16 different conditions, varying in 4 reaction temperatures and 4 reaction times. In contrast, for F8, the deuteration yields were obtained under eight specific conditions: (-78 °C, 4.0 h), (-40 °C, 4.0 h), (-40 °C, 0.5 h), (0 °C, 4.0 h), (0 °C, 2.0 h), (0 °C, 1.0 h), (0 °C, 0.5 h), and (25 °C, 0.5 h). Average yields from multiple experiments under identical conditions are presented in Table 1. While D4 was identifiable, D1, D2, and D3 were not; thus, the products were categorized into D1-D3, D4, and D1-D4 groups. The yield for D1-D4 is the cumulative yield of D1-D3 and D4. These datasets were then utilized for model training. The model's predicted yields were subsequently validated against new experimental data, as detailed in Table 2.

Table 1: Mean yields for training regarding the various reaction temperatures and times.

|  |  |  |
| --- | --- | --- |
| Temperature (°C) | Time (h) | Deuteration yield (%) |
| F7 | F6 | F8 |
| 3,6H-F6 | 3,7H-F6 | D1-D3 | D4 |
| -78.0 | 0.5 | 85.5 | 83.5 | 81.0 | - | - |
| -78.0 | 1.0 | 86.0 | 83.0 | 85.0 | - | - |
| -78.0 | 2.0 | 83.5 | 82.5 | 86.0 | - | - |
| -78.0 | 4.0 | 86.0 | 85.0 | 86.0 | 78.0 |  7.0 |
| -40.0 | 0.5 | 87.0 | 85.5 | 86.0 | 74.0 | 10.0 |
| -40.0 | 1.0 | 86.5 | 84.5 | 85.0 | - | - |
| -40.0 | 2.0 | 83.0 | 83.5 | 88.0 | - | - |
| -40.0 | 4.0 | 76.0 | 84.5 | 89.0 | 65.0 | 16.0 |
|  0.0 | 0.5 | 85.0 | 81.5 | 87.0 | 76.0 | 13.0 |
|  0.0 | 1.0 | 85.0 | 82.5 | 88.0 | 58.0 | 35.0 |
|  0.0 | 2.0 | 46.0 | 83.0 | 89.0 | 50.0 | 46.0 |
|  0.0 | 4.0 | 12.5 | 70.5 | 89.0 | 36.0 | 53.0 |
| 25.0 | 0.5 |  2.0 | 83.5 | 86.0 | 18.0 | 25.0 |
| 25.0 | 1.0 |  0.0 | 77.5 | 88.0 | - | - |
| 25.0 | 2.0 |  0.0 | 63.0 | 86.0 | - | - |
| 25.0 | 4.0 |  0.0 | 54.5 | 83.0 | - | - |

Table 2: Mean yields for validating the various reaction temperatures and times.

|  |  |  |
| --- | --- | --- |
| Temperature (°C) | Time (h) | Deuteration yield of F8 (%) |
| D1-D3 | D4 |
| 25.0 | 1.5 |  8.0 | 15.0 |
| 25.0 | 3.5 |  0.0 |  0.0 |
| -40.0 | 2.0 | 70.0 | 15.0 |
| 15.0 | 3.0 |  9.0 | 24.0 |
| 30.0 | 3.0 |  0.0 |  0.0 |
| 0.0 | 3.0 | 38.0 | 52.0 |
| 40.0 | 3.0 |  0.0 |  0.0 |
| -20.0 | 2.0 | 63.0 | 27.0 |
| 50.0 | 3.0 |  0.0 |  0.0 |

Table 3 displays the predictive performance of the various models for the studied products. To compare with the proposed models, 26 different models from PyCaret (2020) were assessed. These models were evaluated based on the temperature, time, and yields of D1-D3, D4, and D1-D4 from the training datasets. The evaluation criteria focused on the R-squared (R²) value as the objective function, utilizing a 2-fold (k=2) cross-validation algorithm. The best models for D1-D3, D4, and D1-D4 were ‘Light Gradient Boosting Machine: LGBM’, ‘Multi Layer Perceptron Regressor: MLP’, and ‘Gradient Boosting Regressor: GBR’ respectively. The predicted results for F8 were evaluated with the mean absolute error (MAE), root mean squared error (RMSE), and R2 for training and validating datasets. Although the proposed models had less performance than the other models for the training datasets without D1-D3, the proposed models had more performance than the other models for all the validating datasets. Therefore, the proposed models can accurately predict the yields for the non-experimental conditions.

Table 3: Predicting performance of various models for products.

|  |  |  |  |
| --- | --- | --- | --- |
| Product | Model | Training | Validation |
| MAE | RMSE | R2 | MAE | RMSE | R2 |
| D1-D3 | Proposed model | 12.984 | 19.850 | 0.005 | 10.068 | 10.274 | 0.855 |
| LGBM | 20.675 | 22.968 | -0.333 | 33.889 | 36.420 | -0.828 |
| D4 | Proposed model | 15.723 | 21.296 | -0.731 | 13.819 | 17.653 | -0.125 |
| MLP | 10.330 | 13.934 | 0.259 | 35.791 | 41.788 | -5.303 |
| D1-D4 | Proposed model | 14.789 | 19.170 | -0.509 | 15.902 | 18.948 | 0.762 |
| GBR | 5.457 | 10.016 | 0.588 | 39.833 | 48.608 | -0.564 |

The optimized virtual variable ratios were established as (1:2:-2.6:-2.6) for (3,6HF6:3,7HF6:F7:F8) in the case of the D1-D3 yields. For D4, the ratio was set as (2:1:0.01) for (3,7HF6:F7:F8), and for the D1-D4 yields, it was (2:1:1.36) for (3,6HF6:F7:F8). The nearest descriptors by the standard scaled distance to the optimized variables were IC3 or SIC1 for D1-D3, AATSC6s or SM1\_Dzv for D4, and Psi\_i\_0d or GATS6v for D1-D4. Based on these optimized variables and the nearest descriptors, the F8 yields were inferred as follows: the D1-D3 yields were mainly affected by the variety of atoms within 1 or 3 bonds of 3,6HF6, 3,7HF6, F7, and F8. For the D4 yields, the major influences were autocorrelation of the nearest 6 atoms weighted by the intrinsic state or spectral moment weighted by Van-der-Waals volume of 3,7HF6, F7, and F8. Lastly, the D1-D4 yields were primarily determined by the intrinsic state of atoms or autocorrelation of the nearest 6 atoms weighted by Van-der-Waals volume associated with 3,6HF6, F7, and F8. These results indicated that the yields of D1-D3 and D4 were, respectively, affected by atoms within 1 or 3 bonds and atoms within 6 bonds.

Conclusions

This study introduces a novel predictive model for the deuteration yield of polyfluoroperylene, leveraging advanced generative AI techniques. This model stands out in its ability to provide dependable and plausible yield estimates for experimental conditions that have not yet been tested. Key to this model's effectiveness is the virtual variables, which serve as ideal descriptors to represent the relationships between the reactants. Through this approach, we offer a significant tool for streamlining and guiding future experimental research in the field of organic electroluminescent materials.

Acknowledgments

This work was financially supported by a Grant-in-Aid for Transformative Research Areas (A) (Grant Number JP21H05222 and JP21H05218) for Digitalization-driven Transformative Organic Synthesis (Digi-TOS) from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), Japan. Part of this study was conducted at the Institute for Molecular Science and supported by the Advanced Research Infrastructure for Materials and Nanotechnology in Japan (Organic Synthesis DX No. JPMXP1222MS5042 and No. JPMXP1223MS5005) from MEXT, Japan.

References

M. Kondo, H. D. P. Wathsala, K. Ishikawa, D. Yamashita, T. Miyazaki, Y. Ohno, H. Sasai, T. Washio, and S. Takizawa, 2023, Bayesian Optimization-Assisted Screening to Identify Improved Reaction Conditions for Spiro-Dithiolane Synthesis, Molecules, 28, 13, 5180

N. Ohtsuka, H. Ota, S. Sugiura, S. Kakinuma, H. Sugiyama, T. Suzuki, and N. Momiyama, 2023, Perfluorohalogenated Naphthalenes: Synthesis, Crystal Structure, and Intermolecular Interaction, ChemRxiv. Cambridge: Cambridge Open Engage

PyCaret: An open source, low-code machine learning library in Python, 2020, https://www.pycaret.org

S. Saito, M. Nagaoka, T. Nagatomo, and O. Omoto, 1994, Color Variations in Light Emission from Perylene Doped Organic EL Device, The Institute of Image Information and Television Engineers, 18, 2, 65–70

T. J. Struble, T.J., Alvarez, J.C., Brown, S.P., Chytil, M., Cisar, J., DesJarlais, R.L., Engkvist, O., Frank, S.A., Greve, D.R., Griffin, D.J., et al., 2020, Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis, J. Med. Chem., 63, 8667–8682

K. Takeda, N. Ohtsuka, T. Suzuki, and N. Momiyama, 2023, Virtual Variables-Enabled Generation of Datasets for Prediction in Organic Synthesis: Digitalization of Small Molecules and Its Application to Functional Molecule Syntheses, ChemRxiv. Cambridge: Cambridge Open Engage

V. Vapnik, 1995, The Nature of Statistical Learning Theory, Springer, New York

Y. Xiang, D. Y. Sun, W. Fan, and X. G. Gong, 1997, Generalized Simulated Annealing Algorithm and its Application to the Thomson Model, Physics Letters A, 233, 3, 216–220