Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Extraction of physicochemical laws by symbolic regression using a Bayesian information criterion

Naoki Yamane, Kan Hatakeyama-Sato, Yuma Iwasaki & Yasuhiko Igarashi

To cite this article: Naoki Yamane, Kan Hatakeyama-Sato, Yuma Iwasaki & Yasuhiko Igarashi (2024) Extraction of physicochemical laws by symbolic regression using a Bayesian information criterion, Science and Technology of Advanced Materials: Methods, 4:1, 2420658, DOI: 10.1080/27660400.2024.2420658

To link to this article: https://doi.org/10.1080/27660400.2024.2420658

Published online: 12 Nov 2024.

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2024, VOL. 4, NO. 1, 2420658
https://doi.org/10.1080/27660400.2024.2420658

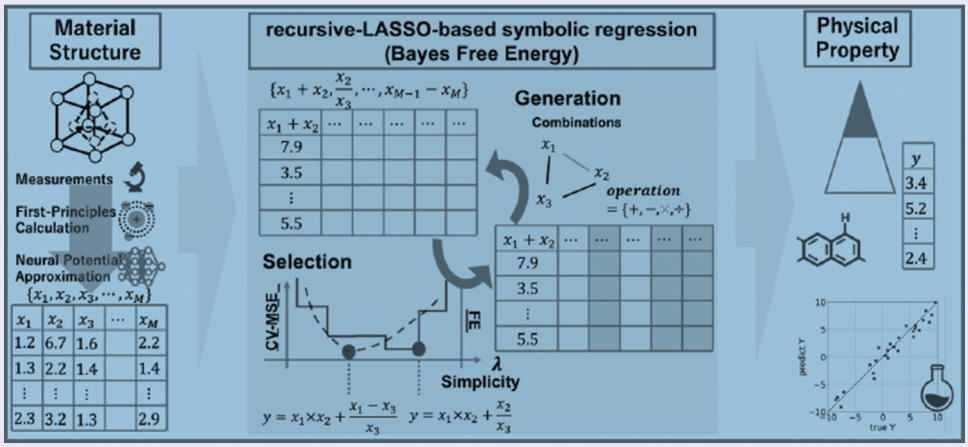
Extraction of physicochemical laws by symbolic regression using a Bayesian information criterion

Naoki Yamane[a], Kan Hatakeyama-Sato[b], Yuma Iwasaki[c] and Yasuhiko Igarashi[d]

[a]Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Ibaraki, Japan; [b]School of Materials and Chemical Technology, Institute of Science Tokyo, Meguro-ku, Tokyo, Japan; [c]Center for Basic Research on Materials (CBRM), National Institute for Materials Science (NIMS), Tsukuba, Ibaraki, Japan; [d]Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki, Japan

ABSTRACT

In the search for new high-performance materials in materials science, especially in polynomial science, it is important to use physicochemical laws linking materials structure and physical properties, and predict the physical properties required for the design. Recently, machine learning (ML) has enabled us to extract patterns from large datasets and construct the data-driven model to predict physical properties. However, ML approach faces challenges such as interpretability and systematic errors of the data-driven model with limited data. Here, we propose a method for extracting an interpretable law from limited data, by combining a symbolic regression method and Bayesian information criterion. We focus on extracting a physicochemical law for the refractive index of polymer materials. The goal is to correct systematic errors and capture physicochemical laws more accurately. Combining explanatory variables from experiments, property calculations, and neural potential approximations, our method involves arithmetic operations on explanatory variables and selection through Bayesian information criterion. The results show that the proposed method is able to correct the results of the neural potential approximation and obtain interpretable and concise expressions for the physicochemical laws linking material structure and physical properties.



ARTICLE HISTORY
Received 11 September 2024
Revised 16 October 2024
Accepted 18 October 2024

KEYWORDS
Symbolic regression; Bayesian information criterion; refractive index; polymer materials; neural potential approximation

IMPACT STATEMENT

This paper introduces a method combining symbolic regression and Bayesian information criterion to extract interpretable physicochemical laws from limited data, improving material property predictions in polymer refractive index analysis.

1. Introduction

Machine learning (ML) has gained attention as an effective approach for extracting patterns from large datasets and constructing predictive models for physical properties. In particular, neural potentials based on deep learning, such as Matlantis developed by PFN [1], have been used to predict general properties like band structures and densities, contributing to the exploration and development of various materials [2]. However, a key challenge moving forward is how to predict more specific properties from limited data since ML-based models face challenges such as the interpretability of data-driven models and systematic errors, particularly when dealing with limited datasets.

CONTACT Yasuhiko Igarashi igayasu1219@cs.tsukuba.ac.jp Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba, Ibaraki 305-8577, Japan

Building on this, one promising approach to overcome the challenges associated with predicting specific properties from limited data is leveraging physicochemical laws. These laws provide a foundational framework that links general properties to key material properties, such as refractiveness [3]. By uncovering such relationships, researchers can bridge the gap between general material characteristics and those properties that are crucial for solving particular problems in materials science. The discovery of these underlying laws not only enhances the interpretability of machine learning models but also reduces systematic errors, ultimately enabling more accurate predictions of specific properties from limited datasets.

This is where techniques like symbolic regression [4–6] come into play. By using symbolic regression, it is possible not only to predict the target variable using explanatory variables but also to obtain the relationships that connect them. The obtained relationships enable us to understand the regression results, thereby enhancing interpretability. In previous studies, methods like those of Cava et al. [7] and Stephens [8], which incorporate genetic algorithms, as well as Udrescu et al.'s [9] approach using properties inherent in physical functions demonstrate the potential of symbolic to reveal unknown relationships from the given data. Iwasaki et al. [10] enhanced this by introducing recursive symbolic regression with cross-validation for optimization, while Weng et al. [11] discovered key relationships in perovskite catalysts using these techniques. However, these symbolic regression methods often suffer from limitations such as insufficient initial features and challenges in exploring synthetic features from limited data, highlighting the need for further improvements.

In this study, we focus on extracting these physicochemical laws using symbolic regression, which allows for the creation of interpretable and concise models from limited data. We propose a symbolic regression method that combines symbolic regression with the Bayesian information criterion to address these challenges. This approach focuses on extracting physicochemical laws related to the refractive index of polymer materials, with the goal of correcting systematic errors in neural potential approximations derived from Matlantis [1]. Our method aims to balance the complexity of nonlinear regression with the need for interpretability. Moreover, we evaluate both the interpolation and extrapolation performance of the model, placing special emphasis on the extrapolation performance, which is the target of this study. By evaluating the model's extrapolation capability, we seek to improve the accuracy of neural potential approximations. This approach not only corrects neural potential predictions but also allows the derivation of concise and interpretable laws that link material structure with physical properties. Our results demonstrate that accurate property predictions can be achieved, even with limited data.

2. Formulation

First, we define the regression problem. Define $\boldsymbol{x_i} \in \mathbb{R}^M$ as the M-dimensional explanatory variable and $y_i \in \mathbb{R}$ as the objective variable. As the dataset, we analyzed data $\mathcal{D} = \{\boldsymbol{x_i}, y_i\}_{i=1}^N$ consisting of N samples. The planning matrix of the explanatory variables was $\boldsymbol{X} \in \mathbb{R}^{\mathbb{N}\times\mathbb{M}}$.

The objective variable was centralized. Explanatory variables were standardized in the selection of symbolic regressions and not in their generation. In the selection of symbolic regression, the explanatory variables were standardized and non-dimensionalized to unify the dimensions of the explanatory variables in the Bayesian linear regression and LASSO sparsification.

In a regression problem with a dataset $\mathcal{D}$ and regression coefficients $\boldsymbol{w} = (w_1, \ldots, w_M)^T$, sparsification by LASSO [12] is based on the assumption that the number of explanatory variables contributing to the objective variable is small, and the penalty term is $\sum_{j=1}^{M} ||w_j||||_1$. The objective is to find the $\boldsymbol{w}$ that minimizes the right side of Equation 1.

$$E(\boldsymbol{w}) = \frac{1}{2n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1^1 \tag{1}$$

$$= \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{M} w_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{M}||w_j||_1^1$$

The parameter λ controls the strength of sparsity. In general, the optimal λ and β from the cross-validation error of Equation 1 when the parameter λ is varied.

Defining symbolic regression based on the definition of Virgolin et al. [13], the function space $\mathcal{F}$ is the set of explanatory variables and operations, and the constant $\beta \in \mathbb{R}$. The function $f : \mathbb{R}^M \to \mathbb{R}$ is composed of a function $f : \mathbb{R}^M \to \mathbb{R}$ that is composited by the original set $\mathcal{P}$ of $\beta \in \mathbb{R}$.

For example, given a source set $\mathcal{P}$, the following functions are contained in the function space.

$$\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}) \in \mathbb{R}^{2\times N}$$

$$\text{operations} = \{+, \times\}$$

$$\mathcal{P} = \{\boldsymbol{X}, \text{operations}, \mathbb{R}\}$$

$$\{f = \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2}, f = \beta_{12}\boldsymbol{x_1} \times \boldsymbol{x_2}\} \in \mathcal{F}$$

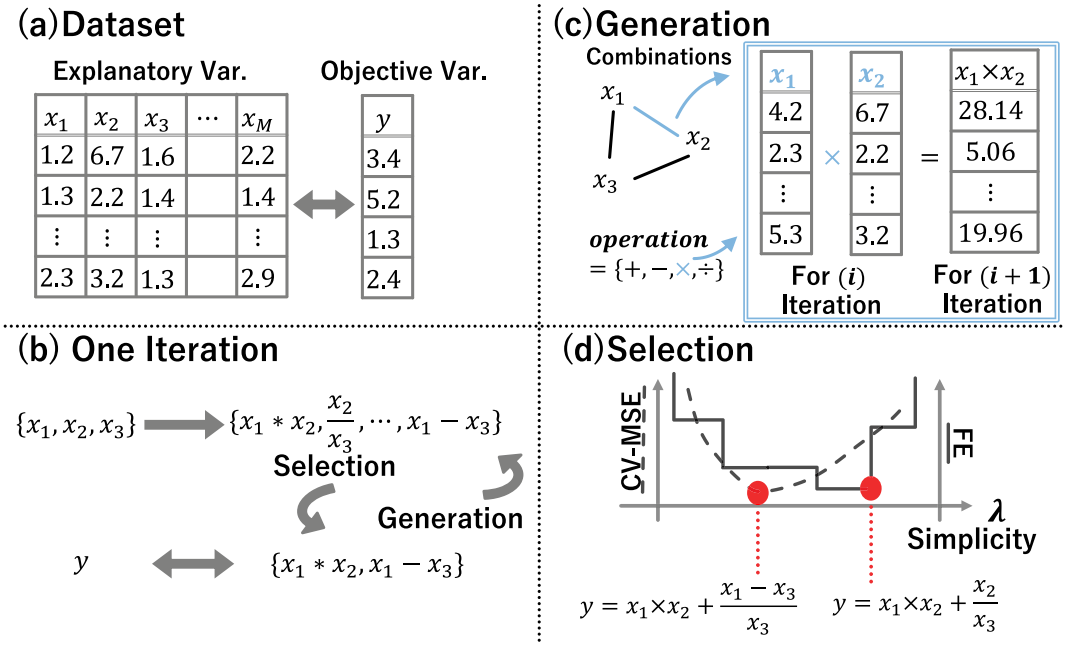


Figure 1. This figure illustrates the process of Recursive Symbolic Regression (RL-Symbolic Reg.), highlighting key steps in the flow. (a) Depicts the format of regression data, consisting of *M* explanatory variables and a target variable. (b) Provides an overview of one iteration in the recursive operation, involving both the exploration of explanatory variables through the generation and pruning of variables through selection. (c) Demonstrates the process of explanatory variable generation, where all combinations of variables from iteration (i) undergo arithmetic operations to introduce new explanatory variables. (d) Highlights the pruning of explanatory variables and showcases that compared to using cross valid error. Bayesian free energy enables the derivation of more concise expressions.

The regression error of a function $f \in \mathcal{F}$ for a dataset $D = (\boldsymbol{x_i}, y_i)_{i=1}^N$, with a hypothesis space $\mathcal{P}$ and an error function l, is measured using the root mean squared error (RMSE).

$$l(f(\boldsymbol{X}), \boldsymbol{Y}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f(\boldsymbol{x_i}) - y_i)^2}$$

In symbolic regression, we seek to find a function $\hat{f} = \arg\min_{f \in F} l(f(\mathbf{X}), \mathbf{Y})$ that minimizes the error. The complexity is defined as the number of initial explanatory variables included in the function f and is denoted as τ.

The process involves generating and selecting functions through L iterations. Figure 1(a) illustrates the data provided for the symbolic regression, while Figure 1(b) shows a single recursive step. Symbolic regression, governed by the primitive set $\mathcal{P}$, constructs complex functions f, and in recursive symbolic regression, these are treated as new explanatory variables. The explanatory variables at the l-th iteration, denoted as $\boldsymbol{X}^{(l)} \in \boldsymbol{R}^{M^{(l)} \times N}$ ($M^{(l)}$ being the number of explanatory variables at the l-th iteration), are defined. The set is represented as $\boldsymbol{X} = (\cdots, \boldsymbol{x_j}, \cdots) = \{\cdots, \boldsymbol{x_j}, \cdots\}$.

The generation phase as shown in Figure 1(c) we perform operations for all combinations of $l-1$ iterations and initial sets of explanatory variables to generate composite features. Then, in the selection phase, as illustrated in Figure 1(d), we determine the set of explanatory variables for each regularization parameter λ, calculate the information content, and select the best set of explanatory variables. $A_m(\tau)$ is a function that controls complexity, and weights are assigned using the following function based on complexity.

Iwasaki's proposed method, Recursive LASSO-based Symbolic Regression with Cross Validation Error (RL-Symbolic Reg. (Cross Validation Error)), used cross-validation error as the information criterion. This error is given by

$$\text{Error} = ||\boldsymbol{y} - \boldsymbol{X}^{(\boldsymbol{i})'}\beta||_2^2 + \lambda \sum_{m=1}^{M^{(i)'}} A_m(\tau)||\beta_m||_1$$

Our use of the Bayesian Free Energy as the information criterion in the selection step of our study deviated from Iwasaki et al.'s choice of cross-validation error. This difference is consistent with the work by Igarashi et al. [14], where the Bayesian Free Energy is expected to generate simpler models with high interpretability. The features generated by symbolic regression are synthetic features, and their validity significantly influences the interpretability of physicochemical laws. Bayesian inference [14,15] is necessary for estimation in order to use symbolic regression effectively. Igarashi et al. [14] have proposed a sparsification method that calculates the Bayesian free energy and cross-validation error for all possible combinations of feature selections and selects the smallest one. By graphically representing the

Bayesian free energy and cross-validation error for all possible combinations, they have demonstrated that the distribution is independents of whether specific are included or not. In addition, they have shown that when the number of explanatory variables is known, LASSO cannot search for better solutions, but when the number is unknown and there are many data, it provides better solutions. Obinata et al. [16] have used Bayesian information criteria to determine the binary inclusion of features and have revealed confident relationships through Bayesian Model Averaging (BMA) of feature means.

Bayesian Free Energy is a Bayesian information criterion that represents the certainty of the target variable given a specific model. Considering a regression problem with a dataset $\mathcal{D}$, assuming the relationship between $\boldsymbol{X_i}$ and $\boldsymbol{y_i}$ is modeled by parameters $\boldsymbol{w} \in \mathbb{R}^M$, basis function $\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}$, and Gaussian noise noise$\sim\mathcal{N}(0, \beta^{-1})$ with $\beta \in \mathbb{R}$, the model $\mathcal{M}$ is assumed with Equation 2.

$$y_i \sim p(y_i|\boldsymbol{x_i}, \boldsymbol{w}, \mathcal{M}) = \mathcal{N}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_i}), \beta^{-1}) \quad (2)$$

At this point, the Bayesian Free Energy is defined as the negative logarithm of the marginal likelihood $p(\boldsymbol{y}|\boldsymbol{X}, \mathcal{M})$, which is expressed in Equation 3.

$$\mathrm{FE} = -\ln p(\boldsymbol{y}|\boldsymbol{X}, \mathcal{M})$$

$$= \frac{1}{2}\left(\beta \boldsymbol{y}^T\boldsymbol{y} - N\ln\beta + N\ln 2\pi + \boldsymbol{m_0}^T\boldsymbol{S_0}^{-1}\boldsymbol{m_0}\right.$$

$$\left. + \ln|\boldsymbol{S_0}| - \boldsymbol{m_N}^T\boldsymbol{S_N}^{-1}\boldsymbol{m_N} - \ln|\boldsymbol{S_N}|\right) \quad (3)$$

In summary, both approaches involve the recursive generation of explanatory variable sets through operations and the selection of explanatory variables through LASSO sparsification. These symbolic regression methods offer an intuitive way to formulate physicochemical phenomena into mathematical expressions.

3. Results and discussion

3.1. Dataset

The experimental dataset used for predicting refractive indices in this study comprised polymer data from the Polymer Properties Database (CROW) [17] and consisted of 47 polymer samples, which represents a limited dataset. Traditionally, the refractive index n is approximated using the Lorentz-Lorenz Equation 4.

Table 1. The explanatory variables for refractive index.

Calculation Method	Number of Variables
Matlantis [1]	1
Gaussian [18]	3
RDKit [19]	50
Group Contribution [20]	15

$$\frac{n^2-1}{n^2+2} = \frac{4\pi}{3}N\alpha \quad (4)$$

Here, N represents the number of molecules per unit volume, and α denotes the polarizability. By rearranging the expression using $N = \frac{6.02\rho}{10MW}$, we obtain the following result.

$$\frac{n^2-1}{n^2+2} = \mathrm{k}\frac{\rho \times \alpha}{\mathrm{MW}}, \mathrm{k} = \frac{4\pi \times 6.02}{3 \times 10} \quad (5)$$

where ρ is the density, α is the polarization factor, MW is the molecular weight, and k is a constant.

Equation 5 shows that the refractive index can be calculated from the density ρ and the polarization factor α. The density ρ and the polarization factor α were therefore used as the objective variables.

The explanatory variables in our predictions consisted of relevant features calculated by various methods, as detailed in Table 1. Matlantis [1] employs an AI-driven molecular simulation approach that leverages the neural potential approximation. Gaussian [18], another simulator, specializes in determining the electronic structure of molecules. RDKit [19], an open-source toolkit for chemoinformatics, generates descriptors crucial for our analysis. Group Contribution [20], which also used RDKit descriptors, predicts physical properties based on molecular structure through a linear model. This method, notably used by Shi et al. [20], forms a significant component of our predictive framework.

The explanatory variables were calculated in the same way as a previously published paper [21]. For details of the meanings and methods of calculating the explanatory variables, see [21].

The density ρ and polarization factor α can be calculated directly using Matlantis and Gaussian, respectively, but because systematic errors occur, symbolic regression was used to extract the physicochemical laws.

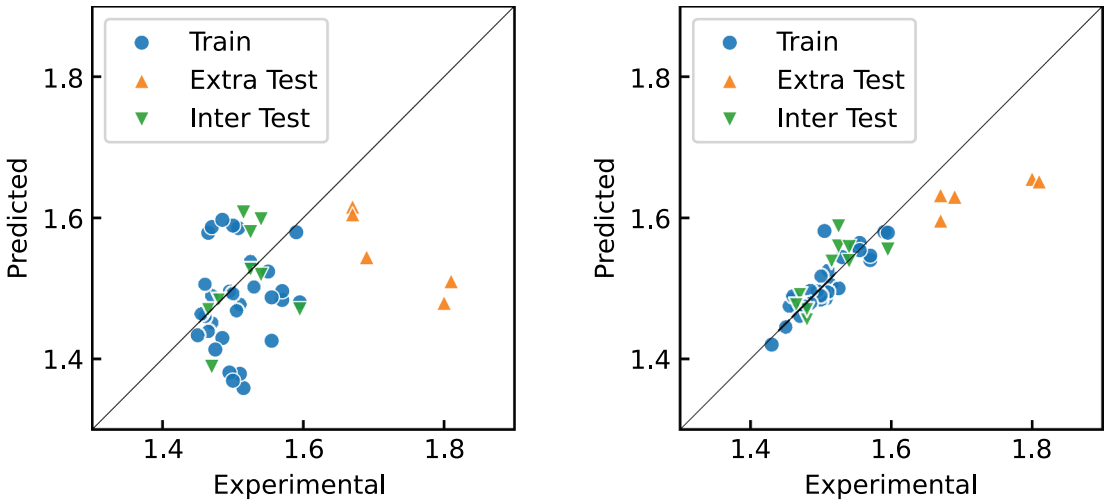
3.2. Experimental setting

The refractive index was then calculated using the predicted ρ and α in Equation 5. The top five refractive indices were used for the test data for extrapolation, and the training data and interpolation predictions were randomly split 8:2.

For comparison methods, we used five techniques: Symbolic Regression (gplearn) [8], LASSO (Cross Valid Error), LASSO (Bayesian Free Energy), RL-Symbolic Reg. (Cross Valid Error), and RL-Symbolic Reg. (Bayesian Free Energy). The parameters for Symbolic Regression (gplearn) were set with 3 iterations, and regression was performed using a genetic algorithm. LASSO (Cross Valid Error) performed the regression and feature selection using cross-validation error with 10 splits.

Table 2. Result of refractive index *n*.undefined. The complexity of the equations and the root mean square error (RMSE) for the training error, interpolation test error, and extrapolation test error.

	Complex	Train	Inter Test	Extra Test
Symbolic Regression (gplearn)	7	8.20×10^{-2}	1.07×10^{-1}	2.10×10^{-1}
LASSO (Cross Valid Error)	8	3.58×10^{-2}	3.02×10^{-2}	1.14×10^{-1}
LASSO (Bayesian Free Energy)	**6**	3.53×10^{-2}	3.07×10^{-2}	1.12×10^{-1}
RL-Symbolic Reg. (Cross Valid Error)	43	2.35×10^{-2}	3.79×10^{-2}	1.34×10^{-1}
RL-Symbolic Reg. (Bayesian Free Energy)	19	$\mathbf{2.01 \times 10^{-2}}$	$\mathbf{2.99 \times 10^{-2}}$	$\mathbf{1.06 \times 10^{-1}}$



(a) Symbolic Regression(gplearn) (b) RL-Symbolic Reg.(Bayesian Free Energy)

Figure 2. Comparison of Predicted Refractive Index n versus Experimental Refractive Index. The horizontal axis represents the experimental values, and the vertical axis represents the predicted values. A good regression is indicated when the data points are distributed along the diagonal line. The results for test data, interpolation tests, and extrapolation tests are distinguished by different markers.

LASSO (Bayesian Free Energy) performed the regression and feature selection with estimated noise parameters of $\rho : \sigma = 1.0 \times 10^{-3}$ and $\alpha : \sigma = 5$ based on Bayesian Free energy. For RL-Symbolic Reg. (Cross Valid Error) and RL-Symbolic Reg. (Bayesian Free Energy), the number of iterations was set to 2, and the set of operators was defined asoperations $= \{\times(\cdot,\cdot), \mathrm{div}(\cdot,\cdot)\}$. The number of splits for cross-validation error was set to 10. The estimated noise parameters for Bayesian Free energy were $\rho : \sigma = 1.0 \times 10^{-3}$ and $\alpha : \sigma = 5$.

RL-Symbolic Reg. (Cross Valid Error) and RL-Symbolic Reg. (Bayesian Free Energy) were used to predict the density,ρ, and polarizability, α. The refractive index was then calculated using the predicted ρ and α in Equation 5. Other methods were employed to directly predict the refractive index.

3.3. Application of symbolic regression

Table 2 shows the complexity and error results as well as the true values, and Figure 2 the comparison of the predicted refractive index n versus the experimental refractive index. The functions obtained by RL-Symbolic Reg. (Bayesian Free Energy) are given in Equations 6 and 7.

$$\rho = 1.37 \times 10^{-1} \times \rho_{\mathrm{RDKit}} \times \rho^2_{\mathrm{Matlantis}} + 7.65 \times 10^{-2} \times \frac{M_{\mathrm{heavy}}}{M_{\mathrm{total}}} + 5.14 \times 10^{-2} \times \frac{M_{\mathrm{heavy}}}{MR_{\mathrm{total}}} \tag{6}$$

$$\alpha = 56.4 \times MR_{\mathrm{total}} + 15.0 \times \frac{T_c}{P_c} - 11.3 \times V_c + 7.67 \times \frac{MR_{\mathrm{total}}}{\rho_{\mathrm{Matlantis}}} + 1.63 \times \left(\frac{f_{\mathrm{amide}}}{MlogP}\right) \times N_{\mathrm{rot}} \tag{7}$$

The variables here are as follows: ρ_{RDKit} is the RDKit density, $\rho_{\mathrm{Matlantis}}$ is the Matlantis density, M_{heavy} is the RDKit HeavyAtomMolWt, M_{total} is the RDKit MolWt, MR_{total} is the RDKit MolMR, T_c is the JR CriticalTemp, P_c is the JR CriticalPress, V_c is the JR CriticalVolume, f_{amide} is the RDKit fr amide, $MlogP$ is the RDKit MolLogP, and N_{rot} is the RDKit NumRotatableBonds.

4. Discussions

From the result with a highly refractive index where the observation noise was unknown, we could deduce the following about hierarchical regression using Bayesian free energy in symbolic regression (RL-Symbolic Reg. (Bayesian Free Energy)):

From Figure 2, which shows the true values and predicted values of the test data, it is apparent that RL-Symbolic Reg. (Bayesian Free Energy) performs regressions that are more accurate in the extrapolation region compared to the results of Symbolic Regression (gplearn).

Let us next consider the relationship expressions for density, ρ, and polarizability, α, obtained using RL-Symbolic Reg. (Bayesian Free Energy).

Equation 6 is the expression for density. The first term involves the neural potential approximation of density (the Matrantis density) multiplied by the molecular weight calculated by RDKit. The second and third terms involve the molecular weight of heavy atom,M_{heavy}, calculated by RDKit. These terms account for the fact that the density increases with the inclusion of heavy atoms and adjusts the density prediction accordingly.

Equation 7 is the expression for polarizability. The second and third terms involve parameters related to the critical temperature, T_c, and the JR critical volume, V_c. The correlation between polarizability and critical temperature has already been reported [22]. These terms represent physicochemical laws that indicate how the critical temperature affects polarizability. The prediction of polarizability, α, is corrected by extracting physicochemical laws solely from data without prior knowledge in a manner similar to how the prediction of density, ρ, is corrected by heavy elements.

When considering the predictive equations, it is important to note their data dependence. If the dataset changes, the coefficients and signs of the parameters may also change, which is an area for future work.

The proposed method was able to achieve an error smaller than the error associated with Symbolic Regression (gplearn). This improvement could be attributed to the regularization effect of using Bayesian free energy with an error model.

It is evident from the results of calculating refractive index based on these predicted densities and polarizabilities that regression can be achieved with concise equations. However, there is an issue with the units in Equation 6 representing the relationship for density, ρ. The calculation of density multiplied by density, $\rho_{\mathrm{RDKit}} \times \rho^2_{\mathrm{Matlantis}}$, cannot be interpreted in terms of units. One future work to resolve this issue is to provide the input unit table proposed by Udrescu et al. [9]. By using this unit table, the method can be improved to ensure that the output relationships are dimensionally consistent by restricting the synthetic features being explored.

5. Conclusions

In this study, we proposed a method that combines symbolic regression with the Bayesian Information Criterion (BIC) to derive interpretable physicochemical laws from limited data. Validation using experimentally measured refractive index data demonstrated that the proposed method can predict outcomes with a simple equation based on interpretable explanatory variables. This underscores the method's ability to balance interpretability and accuracy, providing a promising approach for predicting material properties from small datasets. The predictive equations obtained by the method correct the results of the neural potential approximation; in other words, they connect general variables to key material properties, serving as correction relationships based on the given data. Looking ahead, as methods like Matlantis and neural network potential approximations continue to replace calculations of physicochemical laws with machine learning in material exploration, the applicability of symbolic regression in extracting physicochemical phenomena using intermediate representations from features obtained by different methods is likely to expand and become widely used.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by CREST [JPMJCR21O1] and FOREST Program [Grant Number JPMJFR213V] from the Japan Science and Technology Agency (JST).

ORCID

Kan Hatakeyama-Sato http://orcid.org/0000-0003-1959-5430
Yuma Iwasaki http://orcid.org/0000-0002-7117-277X
Yasuhiko Igarashi http://orcid.org/0000-0003-1042-6657

References

[1] Matlantis. Software as a service style material discovery tool. 2022. Available from: https://matlantis.com/
[2] Takamoto S, Shinagawa C, Motoki D, et al. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. Nat Commun. 2022 May;13(1):2991. doi: 10.1038/s41467-022-30687-9
[3] Muckley ES, Saal JE, Meredig B, et al. Interpretable models for extrapolation in scientific machine

learning. Digit Discov. 2023;2(5):1425–1435. doi: 10.1039/D3DD00082F
[4] Juárez-Smith P, Trujillo L. Integrating local search within neat-gp. In: Proceedings of the 2016 on genetic and evolutionary computation conference companion; Denver, Colorado; 2016. p. 993–996.
[5] Burlacu B, Kronberger G, Kommenda M. Operon c++: an efficient genetic programming framework for symbolic regression. In: Proceedings of the 2020 genetic and evolutionary computation conference companion; 2020. p. 1562–1570. ACM. Available from: https://dl.acm.org/doi/10.1145/3377929.3398099
[6] Kim S, Lu PY, Mukherjee S, et al. Integration of neural network-based symbolic regression in deep learning for scientific discovery. IEEE Trans Neural Netw Learn Syst. 2020;32(9):4166–4177. doi: 10.1109/TNNLS.2020.3017010
[7] La Cava W, Orzechowski P, Burlacu B, et al. Contemporary symbolic regression methods and their relative performance. 2021. Available from: https://arxiv.org/abs/2107.14351
[8] Meurer Aea. Gplearn: genetic programming in Python. 2023. Available from: https://gplearn.readthedocs.io/en/stable/
[9] Udrescu SM, Tegmark M. Ai Feynman: a physics-inspired method for symbolic regression. Sci Adv. 2020;6(16):eaay2631. doi: 10.1126/sciadv.aay2631
[10] Iwasaki Y, Ishida M. Data-driven formulation of natural laws by recursive-lasso-based symbolic regression. arXiv preprint arXiv:210209210. 2021.
[11] Weng B, Song Z, Zhu R, et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. Nat Commun. 2020;11(1):3513. doi: 10.1038/s41467-020-17263-9
[12] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B: Stat Methodol. 1996;58(1):267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x
[13] Virgolin M, Pissis SP. Symbolic regression is np-hard. arXiv preprint arXiv:220701018. 2022.
[14] Igarashi Y, Takenaka H, Nakanishi-Ohno Y, et al. Exhaustive search for sparse variable selection in linear regression. J Phys Soc Jpn. 2018 Apr;87(4):044802. Available from: http://journals.jps.jp/doi/10.7566/JPSJ.87.044802
[15] Nagata K, Kitazono J, Nakajima S, et al. An exhaustive search and stability of sparse estimation for feature selection problem. IPSJ Online Trans. 2015;8(8):25–32. doi: 10.2197/ipsjtrans.8.25
[16] Obinata K, Nakayama T, Ishikawa A, et al. Data integration for multiple alkali metals in predicting coordination energies based on Bayesian inference. Sci Technol Adv Mater: Methods. 2022 Dec;2(1):355–364. Available from: https://www.tandfonline.com/doi/full/10.1080/27660400.2022.2108353
[17] Polymer properties database. 2023. Accessed: [Insert Date of Access]. https://polymerdatabase.com/home.html
[18] Frisch MJ, Trucks GW, Schlegel HB, et al. GaussianËœ16 revision C.01. Wallingford (CT): Gaussian Inc; 2016.
[19] Rdkit. 2022. Available from: https://www.rdkit.org/
[20] Shi C, Borchardt TB. Jrgui: a Python program of Joback and Reid method. ACS Omega. 2017 12;2(12):8682–8688. doi: 10.1021/acsomega.7b01464
[21] Hatakeyama-Sato K, Watanabe S, Yamane N, et al. Using gpt-4 in parameter selection of polymer informatics: improving predictive accuracy amidst data scarcity and 'ugly duckling'dilemma. Digit Discov. 2023;2(5):1548–1557. doi: 10.1039/D3DD00138E
[22] Schmidt JW, Carrillo-Nava E, Moldover MR. Partially halogenated hydrocarbons chfcl cf3, cf3 ch3, cf3 chf chf2, cf3 ch2 cf3, chf2 cf2 ch2f, cf3 ch2 chf2, cf3 o chf2: critical temperature, refractive indices, surface tension and estimates of liquid, vapor and critical densities. Fluid Phase Equilib. 1996;122(1–2):187–206. doi: 10.1016/0378-3812(96)03044-0