Using GPT-4 in Parameter Selection of Materials Informatics: Improving Predictive Accuracy Amidst Data Scarcity and 'Ugly Duckling' Dilemma

Kan Hatakeyama-Sato,*a Seigo Watanabe^b, Naoki Yamane^c, Yasuhiko Igarashi^c, and Kenichi Oyaizu*^b

*K.H.: hatakeyama.k.ac@m.titech.ac.jp

*K.O.: oyaizu@waseda.jp

Abstract

Materials informatics and cheminformatics struggle with data scarcity, hindering the extraction of significant relationships between structures and properties. The "Ugly Duckling" theorem, suggesting the difficulty of data processing without assumptions or prior knowledge, exacerbates this problem. Current methodologies don't entirely bypass this theorem and may lead to decreased accuracy with unfamiliar data. We propose using Open AI Generative Pretrained Transformer 4 (GPT-4) language model for explanatory variable selection, leveraging its extensive knowledge and logical reasoning capabilities to embed domain knowledge in tasks predicting structure-property correlations, such as the refractive index of polymers. This can partially overcome challenges posed by the "Ugly Duckling" theorem and limited data availability.

a. Materials Science and Engineering, School of Materials and Chemical Technology, Tokyo Institute of Technology. Tokyo 152-8552, Japan

b. Department of Applied Chemistry, Waseda University. Tokyo 169-8555, Japan

c. Faculty of Engineering, Information and Systems, University of Tsukuba. Ibaraki 305-8573, Japan

Introduction

Materials informatics and cheminformatics are scientific disciplines aiming to process and derive meaningful chemical and physical insights from correlations between the structures and properties of compounds and materials.¹⁻⁵ One key feature of these approaches is their capacity to rapidly extract statistically significant relationships from constructed databases using data science techniques.⁶ These methodologies have achieved success in fields such as drug discovery and inorganic material exploration.^{4, 5, 7} They enable the efficient extraction of molecules with significant pharmacological activity from a vast pool of candidate substances.^{2, 5, 6} Moreover, in materials science, analyses of large databases have led to the discovery of new luminescent molecules,⁸ ion conductors,^{9, 10} heat conductors,¹¹ and novel alloys.¹²

However, the application of materials informatics in experimental projects faces an apparent problem of data insufficiency. Unlike in the field of biology, where large, standardized databases are available,^{13, 14} such databases are not necessarily present in materials science.¹⁵ This discrepancy can be attributed to a variety of factors.¹⁵ First, the desired material properties span a wide range. Secondly, a unified format for uniquely describing material structures is lacking. Lastly, a standard platform for sharing this data does not always exist. As a result, the number of entries in these databases often remains in the tens.¹⁵⁻²⁰

A well-known theorem in statistics further illustrates the predicament, the "Ugly Duckling" theorem, which suggests that informatics with small-scale data is exceptionally challenging.^{21, 22} This theorem posits that tasks like pattern recognition, classification, and regression are impossible without certain assumptions or prior knowledge. For instance, in the tale of the "Ugly Duckling," the judgment that the black duckling is ugly arises from prior exposure to a large dataset - namely, the common knowledge that typical ducklings are yellow. A person, or an AI, who has only seen two or three ducklings would not be able to judge that the black duckling is ugly.

The "Ugly Duckling" theorem potentially poses a critical problem in materials informatics and cheminformatics.¹⁵ Identifying significant relationships between material structures and properties can become extremely challenging without a large and diverse dataset for these fields. Therefore, the development of standardized, comprehensive databases for materials science is a pressing issue that requires collective action and coordinated effort from the global scientific community.¹⁵

In cheminformatics, the discussion often revolves around the correlation between molecular structure and property. For example, consider the three organic molecules illustrated in Fig. 1a: toluene, cyclohexane, and trimethylamine. It is experimentally known that the boiling point of trimethylamine is lower than that of the other two molecules. Readers with a background in chemistry may intuitively conclude that the lower boiling point of trimethylamine is due to its smaller molecular weight.

However, for readers without a background in chemistry or AI algorithms, identifying the factors that determine boiling points from merely three data points would be challenging. One could posit that the

methyl groups in trimethylamine contribute to its lower boiling point, or perhaps that the presence of the nitrogen atom is responsible. This predicament shares the same logical structure as the dilemma introduced by the "Ugly Duckling" theorem.

Discussing the issue more quantitatively, the dilemma arises from an imbalance between the dimensions of the explanatory variables and the number of data points available for learning. Over years of cheminformatics research, numerous methods have been proposed to describe the characteristics of molecular structures numerically.^{23, 24} Typical molecular descriptors have dimensions in the hundreds, and fingerprints can possess bits in the thousands.^{23, 24} Recent advancements in deep learning for molecular recognition often involve latent vectors of several hundred dimensions.^{7, 25, 26} Consequently, a vector of several hundred dimensions is generally required to characterize a molecular structure.

While there is no general rule determining the minimum ratio between the dimensions of explanatory variables and the number of learning data points, one benchmark suggests that a ratio of over 10:1 is a good starting point.²⁷ Therefore, having at least a few thousand data points is desirable as a general guideline for progressing molecular informatics. Unfortunately, most experimental informatics projects fail to meet this requirement, illustrating a critical challenge in the field.¹⁵



Fig. 1 a) Relationships between chemical structures and their boiling points. b) Steps to extract explanatory variables via domain knowledge and data itself. c) Workflow to conduct prediction tasks via supervised learning.

Several data science methodologies, such as sparse modeling techniques like Lasso and Ridge,²⁸ feature engineerings techniques like Boruta,²⁹ and information criterion (e.g., Akaike criterion),³⁰ have been reported to fit well with databases. These methodologies remove parameters with low relevance to the target variable based on certain statistical rules. Such data science tools are compatible with experimental informatics that struggle with small-scale data, producing notable results.^{10, 16, 20} However, these techniques do not essentially transcend the "Ugly Duckling" theorem, often resulting in decreased accuracy with untrained data.¹⁵

These problems arise from data-oriented feature selection methods, overfitting to the training data, or mismatches between the mathematical algorithms used for dimensionality reduction and the behavior of the target material system. According to the "Ugly Duckling" theorem, the only way to successfully apply machine learning to small-scale data sets is to provide some assumptions or prior knowledge of the regression system. This corresponds to the task of pre-extracting parameters correlated with the predicted properties, based on the domain knowledge of the user of the prediction model — in this case, knowledge of chemistry or materials science.¹⁵

Alternatively, recent deep learning approaches have gained attention, such as transfer learning methods that deploy deep learning models trained on specific structure-property correlations to other material systems.^{9, 11} The deep learning model to be transferred here is an encoder that automatically converts molecular structures into numerical spaces. Thus, specifying a particular trained model for transfer is essentially the same as humans choosing explanatory variables generated by cheminformatics techniques.

These knowledge-oriented variable selection approaches are at the heart of data science (Fig. 1b). However, there has been no effective methodology because variable selection requires a deep understanding of both experimental and data science.¹⁵ This task has been performed by personnel based on their unique intuition and know-how¹⁵ However, these individuals may not have perfect experimental and data science knowledge, and objectively verifying this process in the scientific community has been difficult.

To address these problems, we propose using the large language model GPT-4,³¹ to choose explanatory variables. GPT-4 is a model that possesses vast knowledge, including science, and is capable of logical reasoning. As of the writing of this paper, only two months have passed since its public release, and revolutionary results are being reported one after another by GPT-4 and a large language model (LLM). Examples include the selection of machine learning algorithms,³² predicting structure-property correlations,³³ in-context learning,³⁴ and controlling robotic arms.³⁵ Our approach utilizes the scientific knowledge that GPT-4 possesses for the purpose of embedding domain knowledge in tasks predicting structure-property correlations.

Experimental section

The code related to this study is made available on GitHub (https://github.com/KanHatakeyama/RefractiveIndexGPT).

Dataset preparation

Dataset preparation: We utilized the polymer database available on CROW (https://polymerdatabase.com/), from which we collected data of 44 conventional polymers. The molecular structures of these polymers are shown in Fig. S1. The number 44 represents a value

frequently observed in experimental informatics with a small amount of data. In this study, the unit structure of the polymers was recognized in SMILES format, from which the task of predicting the refractive index was set. The molecular weight and higher-order structures of the polymers were not considered.

Descriptor calculation

Several molecular descriptors were calculated as explanatory variables that can determine the refractive index (Table S 1).

1) Basic molecular descriptors (rdkit_**): We calculated descriptors outputted by the RDKit module (ver. 2022.9.5).

2) Values calculated by DFT (DFT_**): After optimizing the structure using the semi-empirical PM7 level method in Gaussian16, we calculated energy, dipole moment in each axis direction (dipole_), total dipole moment (dipole_Tot), HOMO, LUMO, and molecular polarization at 656 nm with the same method. Calculations were conducted in vacuo.

3) Property values predicted by the group contribution method (JR_**): We calculated predicted values such as boiling point using a reported package (JRgui: A Python Program of Joback and Reid Method).³⁶ 4) Molecular volume estimated by the approximate DFT-MD method (DFTMD_vol): Using a module of the AI molecular simulator Matlantis (v. 3.0.0),³⁷ we estimated the volume occupied by a single molecule. This method packs 20 low molecules into a cell and estimates molecular volume by optimizing the structure (Fig. S2). Although systematic errors accompanied by low molecular approximation were present, the correlation coefficient with the actual measured volume was 0.993, making it a useful value as an explanatory variable.

Regression scheme

A regression task was set with the above parameters as explanatory variables and the actual measured refractive index as the target variable (Fig. 1c). The database was evaluated with 5-fold cross-validation, and mean absolute error (MAE) and root mean squared error (RMSE) were calculated against the validation dataset. As regression models, we selected general algorithms such as Ridge, Lasso, Support Vector Machines (SVM), Gaussian Process Regression (GPR, RBF + white kernel), Random Forest Regression (RFR), and Gradient Boosting Regression (GBR). All were driven on the scikit-learn (1.2.2) module.³⁸ Ridge and Lasso used the RidgeCV and LassoCV classes to optimize the regularization term automatically. Other models, less influenced by hyperparameters, were operated with default settings. During regression, all parameters were standardized using the StandardScaler class of scikit-learn.

Parameter selection as a baseline

The variables used for regression were pre-selected by the following methods as baselines.

All: This is the case where all variables are used (Table S 1).

Random-10: This is the case where 10 randomly selected variables are used (Table S 2).

Random-20: This is the case where 20 randomly selected variables are used (Table S 2).

Boruta: Statistically significant parameters between the explanatory variables and target parameters in each training dataset were selected by the Boruta algorithm with default hyperparameters (v0.3). A Random Forest was chosen as the regressor. In Fig. 3, the descriptor set of Boruta is given as the selected parameters for the train data during first cross-validation step. In the case of Lasso, an entire dataset was introduced to discuss typical descriptors.

Parameter selection by GPT-4

We selected explanatory variables by asking ChatGPT (May 12 Version, GPT-4) for preferred parameters (Fig. 2, see Supplementary Information for the full chat log).³⁹ The question task consists of three parts:

1) Instruct GPT-4 to consider the theoretical formula that determines the refractive index of the molecule. This was to induce reasoning based on physical chemistry.

2) Ask what explanatory variables should be used for polymers. In doing so, we wrote out all the explanatory variables and supplemented the meanings of prefixes such as 'DFT_' and 'rdkit_'.

3) Command GPT-4 to output the list of selected descriptors.

The ChatGPT answers could change each time the question was asked slightly, mainly due to the inclusion of a temperature parameter related to randomness as an internal variable of the LLM. The randomness can be controlled only by using the API. In this study, we adopted the first answer that came up in the interface on ChatGPT.

User prompt 1: Consider theoretical equation of refractive index of molecules 2: Select essential descriptors from below by yourself. Especially, I want to predict refractive index of aliphatic and aromatic polymers. #Followings are list of descriptors : DFT_energy,DFT_dipoleX,DFT_dipoleY, 3: Output the descriptors as list data.
GPT-4 response 1: The refractive index of a material is the ratio of the speed of light in a vacuum to the speed of light in that material. In general, the refractive index of molecules can be theoretically described by the Lorentz-Lorenz equation (also known as the Clausius-Mossotti relation), which is: $(n^2 - 1) / (n^2 + 2) = 4\pi Na/3$ where N is2: Based on the general principles above, the most relevant descriptors from the provided list for predicting the refractive index of aliphatic and aromatic polymers would be: • DFT_energy
 3: Here is the list of selected descriptors: ["DFT_energy", "DFT_HOMO", "DFT_LUMO", "DFT_dipoleTot", "rdkit_MolWt", "rdkit_NumValenceElectrons", "rdkit_NumAromaticRings", "rdkit_NumAliphaticRings", "DFTMD_vol", "rdkit_HeavyAtomMolWt", "rdkit_MaxAbsPartialCharge", "rdkit_MolLogP", "JR_BoilingPoint", "JR_EnthalpyForm"]

Fig. 2 Conversation between the author and GPT-4 to select molecular descriptors. See Supplementary Information for the full text.

Results and discussion

Explanatory parameter extraction by GPT-4

In this study, we employed ChatGPT, the May 12 version, an LLM with knowledge spanning various technical disciplines, including chemistry and programming, as a new approach to molecular descriptor selection. In response to the initial question regarding the theoretical formula to determine the refractive index, GPT-4 accurately identified the Lorentz-Lorenz equation as a key determinant (Fig. 2). As this equation is one of the most common formulas used for predicting the molecular refractive index, ⁴⁰⁻⁴⁴ this verified GPT -4's capability to provide pertinent insights.

Subsequently, we asked GPT-4 to highlight factors essential for predicting the refractive index of polymers based on the theoretical formula. This yielded 14 molecular descriptors, including boiling point, enthalpy of formation, and heavy atom count, as shown in Fig. 3a.



Fig. 3 a) Venn's diagram for the molecular descriptors selected by GPT-4 and Boruta. b) Classification of descriptor sets. Parameters selected by GPT-4 are marked yellow. c) Breakdown of groups in each descriptor set.

Three significant points should be inferred from these responses. First, despite our omission in explaining explanatory parameters, GPT-4 demonstrated the ability to interpret them correctly. For instance, we did not explicitly define the abbreviation "DipoleTot," but GPT-4 replied, "Total dipole moment, directly related to polarizability." This shows that GPT-4 correctly deciphered the meaning of the abbreviation and made an accurate inference on the substantial influence of the dipole moment on polarizability, a critical factor in the Lorentz-Lorenz equation.

The second advantage of this approach lies in the physical-chemical relevance of the chosen parameters. According to the Lorentz-Lorenz equation, molecular density (volume) and polarizability are vital in determining the refractive index. These parameters correspond to DFT_alpha656nm and

DFTMD_vol in our descriptors list. Although DFT_alpha656nm was not selected by GPT-4, the DipoleTot, which can serve as a substitute, was chosen. Furthermore, as we expected, DFTMD_vol was set. In addition to these two parameters, GPT-4 also selected descriptors like boiling point and enthalpy of formation, which could be seen as corrective factors. These parameters strongly correlate with the refractive index,⁴⁰⁻⁴⁴ and their selection demonstrates the chemically sound judgment made by GPT-4.

The third point of significance lies in the interpretability of the chosen parameters. All the selected variables possessed straightforward physical-chemical meanings and were accompanied by concise explanatory notes from GPT-4. The series of inferential processes carried out by GPT-4 seemed to agree with the decision-making of an experienced researcher well-versed in chemistry and informatics.⁴⁰⁻⁴⁴ This suggests that embedding domain knowledge could be delegated to an LLM. Furthermore, the fact that this process of descriptor selection was accomplished in significantly less time than a human (in a matter of seconds) is intriguing. A human researcher, requiring ample time for reading and interpreting the descriptors, would need tens of minutes or more. This finding hints at the potential for significant efficiency gains by integrating LLMs in similar scientific workflows.

Data-oriented explanatory parameter extraction by Boruta and Lasso

Boruta algorithm extracts statistically significant explanatory variables through repetitive testing procedures. For a given dataset, Boruta identified the following six explanatory variables: EState_VSA7, FractionCSP3, GibbsEnergy, HOMO, SMR_VSA10, VSA_EState6 (Fig. 3a). The noticeable differences between the variable selection results of GPT-4 and Boruta lay in a) the number of selected variables, and b) the nature of the variables themselves.

GPT-4 proposed 14 parameters as candidate explanatory variables, whereas Boruta only suggested six. Given a dataset size of 44, this outcome could be deemed reasonable. However, parameters statistically significant in the training dataset do not necessarily remain effective in unfamiliar datasets. This assumption holds only when the dataset size is sufficiently large and the quality of training and validation (or testing) datasets is comparable. For small-scale data, as in our case, the applicability of such a hypothesis is dubious. Another issue with Boruta is the interpretability of the proposed explanatory variables. Although the suggested parameters, such as EState_VSA7, SMR_VSA10, and VSA_EState6, might help predict the refractive index, they are highly specialized parameters that no one outside of cheminformatics experts may fully interpret.

Lasso Regression, a widespread technique in statistical modeling, has also been utilized to extract a series of parameters (Fig. S3). The list of extracted parameters includes the following: BalabanJ, CriticalTemp, dipoleX, dipoleY, EState_VSA7, EState_VSA8, fr_Al_COO, fr_COO, fr_COO2, fr_NH2, fr_priamide, FractionCSP3, GibbsEnergy, LUMO, qed, SlogP_VSA12, SlogP_VSA8, SMR_VSA4, SMR_VSA7, and VSA_EState3. These parameters cover a broad spectrum, ranging from conventional descriptors like critical temperature (CriticalTemp) to more specific ones such as the

functional group counts (e.g., fr_NH2, fr_priamide). Approximately half of these parameters might not be immediately familiar to experimental chemists.

Comparison of parameter selection trends

In order to enhance our understanding of the trends in the parameters extracted by each algorithm, we classified each parameter into one of four groups (Fig. 3b). Group 1 encompasses parameters that determine refractive index based on the Lorentz-Lorenz equation, specifically, polarizability (α) and volume (*V*). Group 2 consists of physical properties that are relatively familiar to chemists, such as boiling points, HOMO, and LUMO energies. Group 3 refers to typical molecular descriptors, such as molecular weight and MaxAbsPartialCharge. Group 4 contains geometric descriptors, like counts of specific functional groups.

Each group can be understood as forming a hierarchical structure. Lower-numbered groups (i.e., Groups 1 and 2) are seen as dependent variables of higher-numbered groups. Although there are no golden rules for selecting parameters for regression, as a rule of thumb, smaller group number parameters are more likely to generalize well. For instance, a descriptor counting the number of $-NH_2$ bonds in a molecule (Group 4) does not generalize well to bonds with similar properties like >NH. If one wants to discuss molecular polarity, one should use higher-order parameters such as dipole (Group 2), which are determined as dependent variables of the various functional groups. Therefore, an effective variable selection strategy for small datasets is to choose parameters likely to generalize well (i.e., those from smaller group numbers) as the basis, then add lower-order variables as correction terms.

Interestingly, about 60% of the explanatory variables selected by GPT-4 belonged to Groups 1 and 2, indicating a justifiable parameter selection from a physical chemistry perspective (Fig. 3c). Conversely, the proportions for Boruta and Lasso were approximately 30% and 25%, respectively. In other words, from a physical chemistry perspective, these data-oriented methods chose parameters that are not necessarily easy to generalize in over 70% of cases, which may have contributed to a decline in the aftermentioned performance in cross-validation. Such data-oriented approaches can only deal with raw numerical data, without considering the meaning of the given explanatory variables. They fail to consider the hierarchical nature of the parameters, an issue directly connected to the "Ugly Duckling" theorem. If data-oriented approaches want to account for hierarchical relationships between parameters, they must prepare a suitable scale database. However, this is often impractical in experimental informatics due to the high data acquisition cost, thus leaving the dilemma unresolved.

On the other hand, GPT-4, having been trained on a vast amount of scientific data, can make feature selections that consider the context of physical chemistry and the hierarchy of the parameters. We can call this a knowledge-oriented feature selection. We performed prompts to infer their causal relationships to confirm whether GPT-4 truly understands the hierarchical relationships between variables (Fig. 4, see Supplementary Information for the actual conversation). The model postulated that the refractive index is a function of a molecule's energy, HOMO, LUMO, dipole moment, and

volume. It also suggested that these parameters are functions of lower-order variables such as the number of valence electrons and molecular weight. There were some instances where the model made incorrect physical-chemical interpretations, such as boiling point being an independent variable for the number of valence electrons. The cause seems to be either that GPT-4's inferential ability is still insufficient, that it lacks sufficient chemical knowledge, or possibly both. Regardless, the relationships between variables were generally captured correctly. Combining this with data-oriented methods is expected to achieve higher regression performance.



Fig. 4 Suggested relationships of parameters by GPT-4.

Regression results

Fig. 5a illustrates the performance provided by each descriptor set (All, Random-10, 20, Boruta, GPT-4) generated by the respective algorithms during a 5-fold cross-validation task (see Table 1 and Table S 3 for actual values). Overall, the combination of GPT-4 and Ridge (MAE = 0.0229) offered the best regression performance, followed by Lasso (0.0236). RMSE followed a similar trend with MAE, indicating that outliers did not affect results significantly. The next best performance was achieved under the condition of using all variables with Lasso (0.0272). In GPT -4's descriptor set, only one parameter (MaxAbsPartialCharge) had a coefficient reduced to zero by Lasso regression, indicating that

GPT-4 selected instrumental variables for defining the refractive index from both physical chemistry and statistical perspectives (Fig. S3b). These results imply that if a crucial variable selection is conducted in advance by applying physical chemistry knowledge, the performance of the regression model can be improved.

In other words, it is demonstrated that tasks traditionally considered unique to human expertise, such as embedding domain knowledge, can be delegated to LLMs. To our knowledge, there are scarcely any reported material research examples of assigning such high-level abstract intellectual work in data science to AI.

Another implication of these findings is that even algorithms that automatically perform variable selection, like sparse modeling, cannot ideally determine parameters solely based on the given dataset. This result entirely agrees with the "Ugly Duckling" theorem. Boruta, another critical baseline, provided a relatively good predictive performance on average (MAE = ca. 0.03) but failed to reach the maximum performance of GPT-4. This reveals the limitations of variable selection procedures that solely rely on the given dataset, much like sparse modeling.



Fig. 5 Summary of regression results. a) MAE versus descriptor for cross-validation. b) Relationship between experimental and predicted refractive index for an example of Ridge regressor with "All" descriptor set. c) Boruta set. d) GPT-4 set. Typical validation results are shown for b) to d).

Descriptor	Model	MAE	RMSE
GPT-4	Ridge	0.0229	0.0310
GPT-4	Lasso	0.0236	0.0320
All	Lasso	0.0272	0.0334
Boruta	GPR	0.0276	0.0339
Boruta	Ridge	0.0279	0.0345
All	GBR	0.0287	0.0374
All	Ridge	0.0291	0.0373
Boruta	RFR	0.0295	0.0391
All	RFR	0.0297	0.0394
Boruta	GBR	0.0304	0.0384

Table 1 Best ten regression conditions for MAE.

The worst predictive performance occurred when descriptors were randomly selected, likely due to the omission of crucial variables affecting the refractive index. The performance of GPR improved with the use of Boruta and GPT-4. Since GPR lacks an internal system for variable selection, the model probably failed with the 'All' option due to an excess of explanatory variables. SVR generally delivered poor predictive performance, possibly due to an overabundance of model degrees of freedom relative to the available data for learning. RFR and GBR models, which typically have high predictive performance, failed to match the conditions of GPT-4 combined with linear models, even when combined with Boruta, for the limited data set in this study. Given the nature of these decision tree-based algorithms, it is plausible that they struggled to regress flexibly on small datasets.

To observe actual prediction plots, examples of regression results performed during cross-validation using the Ridge model with several descriptor sets are presented in Fig. 5b-d. When all descriptors were used, there was a variance in the overall correlation with the actual values. In the case of using Boruta, the predictions were biased toward the average value. This could be due to Boruta selecting explanatory variables using a nonlinear mechanism, which may not have been appropriate for ridge regression. Conversely, GPT-4 reduced variance and bias compared to the previous two methods, supporting the efficacy of variable selection based on domain knowledge.

Conclusions

In this study, we demonstrated that the preselection of explanatory variables, a task historically performed by human scientists through the application of domain knowledge, can be delegated to large language models like GPT-4. Large language models can quickly and objectively verify and delegate such tasks. For example, we set a task to predict the refractive index from molecular structure using a small dataset of approximately 40 polymer records. Instead of providing an actual dataset, we asked GPT-4, which 'knows' physical chemistry, to select variables that could influence the refractive index. This approach revealed that superior predictive performance could be achieved compared to algorithms like Boruta and sparse modeling that perform variable selection based on given only numerical datasets. This success indicates the importance of utilizing domain knowledge, especially for small datasets, aligning with the "Ugly Duckling Theorem."

Future work needs to incorporate chemical and material information for more specialized targets. Since GPT-4 only has general physical chemistry knowledge, mechanisms for learning through retrieving recent literature or fine-tuning are necessary. Building more advanced large language models could also automate machine learning tasks by suggesting descriptors to calculate from scratch or designing regression models. Alternatively, combining with models like symbolic regression that align well with theoretical research could provide interpretability and superior predictive performance. The linguistic selection process must also be revealed objectively by scientific approaches. By continuing to investigate these specific cases, we aim to uncover the extent to which large language models can serve as a potent tool in various chemical and material sciences.

Author Contributions

K.H. wrote the code, designed the experiments, and wrote the manuscript. Other authors participated in the discussion.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was partially supported by Grants-in-Aid for Scientific Research (Nos. 21H04695, 21H02017, 22KJ2927, and 22H04623) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, by the JST FOREST Program (Grant Number JPMJFR213V). The manuscript was drafted using GPT-4.

References

1. L. M. Antunes, R. Grau-Crespo and K. T. Butler, Npj Comput. Mater., 2022, 8, 44.

- 2. W. Chaikittisilp, Y. Yamauchi and K. Ariga, *Adv. Mater.*, 2022, **34**, e2107212.
- 3. J. Hu, S. Stefanov, Y. Song, S. S. Omee, S.-Y. Louis, E. M. D. Siriwardane, Y. Zhao and L. Wei, *Npj Comput. Mater.*, 2022, **8**, 65.
- 4. L. Shen, J. Zhou, T. Yang, M. Yang and Y. P. Feng, Acc. Mater. Res., 2022, 3, 572-583.
- 5. Y. Chen and J. Kirchmair, *Mol. Inform.*, 2020, **39**, e2000171.
- E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, Baskin,
 II, A. Varnek, A. Roitberg, O. Isayev, S. Curtarolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A.
 Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chem. Soc. Rev.*, 2020, 49, 3525-3564.
- 7. D. Jiang, Z. Wu, C. Y. Hsieh, G. Chen, B. Liao, Z. Wang, C. Shen, D. Cao, J. Wu and T. Hou, *J. Cheminform.*, 2021, **13**, 12.
- T. C. Wu, A. Aguilar-Granda, K. Hotta, S. A. Yazdani, R. Pollice, J. Vestfrid, H. Hao, C. Lavigne, M. Seifrid, N. Angello, F. Bencheikh, J. E. Hein, M. Burke, C. Adachi and A. Aspuru-Guzik, *Adv. Mater.*, 2023, 35, e2207070.
- 9. K. Hatakeyama-Sato, T. Tezuka, M. Umeki and K. Oyaizu, J. Am. Chem. Soc., 2020, 142, 3301-3305.
- K. Hatakeyama-Sato, M. Umeki, H. Adachi, N. Kuwata, G. Hasegawa and K. Oyaizu, *Npj Comput. Mater.*, 2022, 8, 170.
- S. Wu, Y. Kondo, M.-a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu,
 J. Shiomi, C. Schick, J. Morikawa and R. Yoshida, *Npj Comput. Mater.*, 2019, 5, 66.
- 12. J. M. Rickman, H. M. Chan, M. P. Harmer, J. A. Smeltzer, C. J. Marvel, A. Roy and G. Balasubramanian, *Nat. Commun.*, 2019, **10**, 2618.
- 13. L. Y. Geer, A. Marchler-Bauer, R. C. Geer, L. Han, J. He, S. He, C. Liu, W. Shi and S. H. Bryant, *Nucleic Acids Res.*, 2010, **38**, D492-496.
- J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583-589.
- 15. K. Hatakeyama-Sato, *Polym. J.*, 2022, **55**, 117-131.
- T. Komura, K. Sakano, Y. Igarashi, H. Numazawa, H. Imai and Y. Oaki, ACS Appl. Energy Mater., 2022, 5, 8990-8998.
- 17. Y. Zhang and C. Ling, *Npj Comput. Mater.*, 2018, **4**, 25.
- S. Pruksawan, G. Lambard, S. Samitsu, K. Sodeyama and M. Naito, *Sci. Technol. Adv. Mater.*, 2019, 20, 1010-1021.
- 19. G. Lambard and E. Gracheva, Mach.Llearn.: Sci. Technol., 2020, 1, 025004.
- 20. R. Mizuguchi, Y. Igarashi, H. Imai and Y. Oaki, *Nanoscale*, 2021, 13, 3853-3859.
- 21. S. Watanabe, *Knowing and guessing; a quantitative study of inference and information* New York: Wiley, 1969.

- 22. Z. Wang, S. Chen, J. Liu and D. Zhang, IEEE Trans Neural Netw, 2008, 19, 758-769.
- 23. H. Moriwaki, Y. S. Tian, N. Kawashita and T. Takagi, J. Cheminform., 2018, 10, 4.
- 24. A. Capecchi, D. Probst and J. L. Reymond, J. Cheminform., 2020, 12, 43.
- 25. J. Park, Y. Shim, F. Lee, A. Rammohan, S. Goyal, M. Shim, C. Jeong and D. S. Kim, *ACS Polym. Au.*, 2022, **2**, 213-222.
- 26. O. Wieder, S. Kohlbacher, M. Kuenemann, A. Garon, P. Ducrot, T. Seidel and T. Langer, *Drug Discov Today Technol*, 2020, **37**, 1-12.
- 27. P. Peduzzi, J. Concato, A. R. Feinstein and T. R. Holford, J. Clin. Epidemiol., 1995, 48, 1503-1510.
- 28. J. Mairal, F. Bach and J. Ponce, 2014, arXiv:1411.3230.
- 29. M. Kursa and W. Rudnicki, J. Stat. Softw., 2010, 36, 1-13.
- 30. P. Stoica and Y. Selen, *IEEE Signal Process. Mag.*, 2004, 21, 36-47.
- 31. OpenAI, GPT-4 Technical Report, <u>https://cdn.openai.com/papers/gpt-4.pdf</u>, 2023.
- 32. S. Zhang, C. Gong, L. Wu, X. Liu and M. Zhou, 2023, arXiv:2305.02499.
- 33. A. M. Bran, S. Cox, A. D. White and P. Schwaller, 2023, arXiv:2304.05376.
- 34. M. C. Ramos, S. S. Michtavy, M. D. Porosoff and W. A. D., 2023, arXiv:2304.05341.
- M. Skreta, N. Yoshikawa, S. Arellano-Rubach, Z. Ji, L. B. Kristensen, K. Darvish, A. Aspuru-Guzik,
 F. Shkurti and A. Garg, 2023, arXiv:2303.14100.
- 36. C. Shi and T. B. Borchardt, ACS Omega, 2017, 2, 8682-8688.
- S. Takamoto, C. Shinagawa, D. Motoki, K. Nakago, W. Li, I. Kurata, T. Watanabe, Y. Yayama, H. Iriguchi, Y. Asano, T. Onodera, T. Ishii, T. Kudo, H. Ono, R. Sawada, R. Ishitani, M. Ong, T. Yamaguchi, T. Kataoka, A. Hayashi, N. Charoenphakdee and T. Ibuka, *Nat. Commun.*, 2022, 13, 2991.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer,
 R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, J. Mach. Learn. Res., 2011, 12, 2825-2830.
- 39. https://chat.openai.com/
- 40. F. Gharagheizi, P. Ilani-Kashkouli, A. Kamari, A. H. Mohammadi and D. Ramjugernath, *J. Chem. Eng. Data*, 2014, **59**, 1930-1943.
- 41. P. R. Duchowicz, S. E. Fioressi, D. E. Bacelo, L. M. Saavedra, A. P. Toropova and A. A. Toropov, *Chemom. Intell. Lab. Syst.*, 2015, **140**, 86-91.
- 42. V. Venkatraman and B. K. Alsberg, *Polymers*, 2018, **10**, 103.
- 43. J. P. Lightstone, L. Chen, C. Kim, R. Batra and R. Ramprasad, J. Appl. Phys., 2020, 127, 215105.
- 44. S. A. Schustik, F. Cravero, I. Ponzoni and M. F. Díaz, Comput. Mater. Sci., 2021, 194, 110460.

Supplementary Information

Using GPT-4 in Parameter Selection of Materials Informatics: Improving Predictive Accuracy Amidst Data Scarcity and 'Ugly Duckling' Dilemma

Kan Hatakeyama-Sato,* Seigo Watanabe, Naoki Yamanec, Yasuhiko Igarashi, and Kenichi Oyaizu*



Fig. S1 Chemical structures recorded in the polymer database.

DFT_energy	rdkit_MaxEStateIndex	rdkit_Kappa1	rdkit_EState_VSA8	rdkit_fr_N_O	rdkit_fr_phos_ester
DFT_dipoleX	rdkit_MinEStateIndex	rdkit_Kappa2	rdkit_EState_VSA9	rdkit_fr_Ndealkylation1	rdkit_fr_piperdine
DFT_dipoleY	rdkit_MaxAbsEStateIndex	rdkit_Kappa3	rdkit_VSA_EState1	rdkit_fr_Ndealkylation2	rdkit_fr_piperzine
DFT_dipoleZ	rdkit_MinAbsEStateIndex	rdkit_LabuteASA	rdkit_VSA_EState10	rdkit_fr_Nhpyrrole	rdkit_fr_priamide
DFT_dipoleTot	rdkit_qed	rdkit_PEOE_VSA1	rdkit_VSA_EState2	rdkit_fr_SH	rdkit_fr_prisulfonamd
DFT_HOMO	rdkit_MolWt	rdkit_PEOE_VSA10	rdkit_VSA_EState3	rdkit_fr_aldehyde	rdkit_fr_pyridine

Table S 1 List of molecular descriptors employed in this study.

DFT_LUMO	rdkit_HeavyAtomMolWt	rdkit_PEOE_VSA11	rdkit_VSA_EState4	rdkit_fr_alkyl_carbamate	rdkit_fr_quatN
DFT_alpha656nm	rdkit_ExactMolWt	rdkit_PEOE_VSA12	rdkit_VSA_EState5	rdkit_fr_alkyl_halide rdkit_fr_sulfide	
JR_BoilingPoint	rdkit_NumValenceElectrons	rdkit_PEOE_VSA13	rdkit_VSA_EState6	rdkit_fr_allylic_oxid rdkit_fr_sulfonam	
JR_MeltingPoint	rdkit_NumRadicalElectrons	rdkit_PEOE_VSA14	rdkit_VSA_EState7	rdkit_fr_amide rdkit_fr_sulfone	
JR_CriticalTemp	rdkit_MaxPartialCharge	rdkit_PEOE_VSA2	rdkit_VSA_EState8	rdkit_fr_amidine	rdkit_fr_term_acetylene
JR_CriticalPress	rdkit_MinPartialCharge	rdkit_PEOE_VSA3	rdkit_VSA_EState9	rdkit_fr_aniline	rdkit_fr_tetrazole
JR_CriticalVolume	rdkit_MaxAbsPartialCharge	rdkit_PEOE_VSA4	rdkit_FractionCSP3	rdkit_fr_aryl_methyl	rdkit_fr_thiazole
JR_EnthalpyForm	rdkit_MinAbsPartialCharge	rdkit_PEOE_VSA5	rdkit_HeavyAtomCount	rdkit_fr_azide	rdkit_fr_thiocyan
JR_GibbsEnergy	rdkit_FpDensityMorgan1	rdkit_PEOE_VSA6	rdkit_NHOHCount	rdkit_fr_azo	rdkit_fr_thiophene
JR_HeatCapacity	rdkit_FpDensityMorgan2	rdkit_PEOE_VSA7	rdkit_NOCount	rdkit_fr_barbitur	rdkit_fr_unbrch_alkane
JR_EnthalpyVap	rdkit_FpDensityMorgan3	rdkit_PEOE_VSA8	rdkit_NumAliphaticCarbocycles	rdkit_fr_benzene	rdkit_fr_urea
JR_EnthalpyFus	rdkit_BCUT2D_MWHI	rdkit_PEOE_VSA9	rdkit_NumAliphaticHeterocycles	rdkit_fr_ben	zodiazepine
JR_LiquidVisco	rdkit_BCUT2D_MWLOW	rdkit_SMR_VSA1	rdkit_NumAliphaticRings	rdkit_fr_	bicyclic
JR_CrystalSolub_1	rdkit_BCUT2D_CHGHI	rdkit_SMR_VSA10	rdkit_NumAromaticCarbocycles	rdkit_fr	_diazo
JR_CrystalSolub_2	rdkit_BCUT2D_CHGLO	rdkit_SMR_VSA2	rdkit_NumAromaticHeterocycles	rdkit_fr_dihy	dropyridine
JR_AmorphSolub_1	rdkit_BCUT2D_LOGPHI	rdkit_SMR_VSA3	rdkit_NumAromaticRings	rdkit_fr_epoxide	
JR_AmorphSolub_2	rdkit_BCUT2D_LOGPLOW	rdkit_SMR_VSA4	rdkit_NumHAcceptors	rdkit_f	r_ester
DFTMD_vol	rdkit_BCUT2D_MRHI	rdkit_SMR_VSA5	rdkit_NumHDonors	rdkit_fr_ether	
rdkit_vol	rdkit_BCUT2D_MRLOW	rdkit_SMR_VSA6	rdkit_NumHeteroatoms	rdkit_fr_furan	
	rdkit_BalabanJ	rdkit_SMR_VSA7	rdkit_NumRotatableBonds	rdkit_fr_guanido	
	rdkit_BertzCT	rdkit_SMR_VSA8	rdkit_NumSaturatedCarbocycles	rdkit_fr_halogen	
	rdkit_Chi0	rdkit_SMR_VSA9	rdkit_NumSaturatedHeterocycles	rdkit_fr_hdrzine	
	rdkit_Chi0n	rdkit_SlogP_VSA1	rdkit_NumSaturatedRings	rdkit_fr_hdrzone	
	rdkit_Chi0v	rdkit_SlogP_VSA10	rdkit_RingCount	rdkit_fr_imidazole	
	rdkit_Chi1	rdkit_SlogP_VSA11	rdkit_MolLogP	rdkit_fr_imide	
	rdkit_Chi1n	rdkit_SlogP_VSA12	rdkit_MolMR	rdkit_fr_	isocyan
	rdkit_Chi1v	rdkit_SlogP_VSA2	rdkit_fr_Al_COO	rdkit_fr_isothiocyan	
	rdkit_Chi2n	rdkit_SlogP_VSA3	rdkit_fr_Al_OH	rdkit_fr_ketone	
	rdkit_Chi2v	rdkit_SlogP_VSA4	rdkit_fr_Al_OH_noTert	rdkit_fr_ketone_Topliss	
	rdkit_Chi3n	rdkit_SlogP_VSA5	rdkit_fr_ArN	rdkit_fr_lactam	
	rdkit_Chi3v	rdkit_SlogP_VSA6	rdkit_fr_Ar_COO	rdkit_fr_lactone	
	rdkit_Chi4n	rdkit_SlogP_VSA7	rdkit_fr_Ar_N	rdkit_fr_methoxy	
	rdkit_Chi4v	rdkit_SlogP_VSA8	rdkit_fr_Ar_NH	rdkit_fr_morpholine	
	rdkit_HallKierAlpha	rdkit_SlogP_VSA9	rdkit_fr_Ar_OH	rdkit_fr_nitrile	
	rdkit_Ipc	rdkit_TPSA	rdkit_fr_COO	rdkit_f	r_nitro
		rdkit_EState_VSA1	rdkit_fr_COO2	rdkit_fr_n	itro_arom
		rdkit_EState_VSA10	rdkit_fr_C_O	rdkit_fr_nitro_a	arom_nonortho

	rdkit_EState_VSA11	rdkit_fr_C_O_noCOO	rdkit_fr_nitroso
	rdkit_EState_VSA2	rdkit_fr_C_S	rdkit_fr_oxazole
	rdkit_EState_VSA3	rdkit_fr_HOCCN	rdkit_fr_oxime
	rdkit_EState_VSA4	rdkit_fr_Imine	rdkit_fr_para_hydroxylation
	rdkit_EState_VSA5	rdkit_fr_NH0	rdkit_fr_phenol
	rdkit_EState_VSA6	rdkit_fr_NH1	rdkit_fr_phenol_noOrthoHbond
	rdkit_EState_VSA7	rdkit_fr_NH2	rdkit_fr_phos_acid

 Table S 2 Molecular descriptor sets for Random-10 and 20.

Random-10	Random-20
rdkit_fr_urea	rdkit_fr_phos_ester
rdkit_EState_VSA6	rdkit_fr_thiazole
rdkit_fr_oxazole	rdkit_SMR_VSA9
JR_LiquidVisco	rdkit_NumHAcceptors
rdkit_VSA_EState5	rdkit_EState_VSA1
rdkit_MinEStateIndex	rdkit_fr_NH2
rdkit_PEOE_VSA8	JR_CrystalSolub_1
rdkit_MolLogP	DFTMD_vol
rdkit_NumHeteroatoms	rdkit_PEOE_VSA10
rdkit_VSA_EState10	rdkit_RingCount
	rdkit_BertzCT
	rdkit_SMR_VSA4
	rdkit_fr_nitroso
	rdkit_FpDensityMorgan2
	rdkit_fr_allylic_oxid
	rdkit_fr_prisulfonamd
	rdkit_PEOE_VSA6
	rdkit_fr_Ar_COO
	rdkit_fr_imide
	rdkit_fr_sulfide



Fig. S2 Relationship between calculated and experimental molecular volume (angstrom unit). Calculations were conducted by Matlantis. Experimental values were derived from actual density and molecular weight.

Chat log between the authors and GPT-4.

(1) Parameter selection

Question by Author

Tasks

1: consider theoretical equation of refractive index of molecules

2: select essential descriptors from below by yourself

Especially, I want to predict refractive index of aliphatic and aromatic polymers.

#Followings are list of descriptors

DFT_energy,DFT_dipoleX,DFT_dipoleY,DFT_dipoleZ,DFT_dipoleTot,DFT_HOMO,DFT_LUMO,DFT_alp ha656nm,rdkit_vol,rdkit_MaxEStateIndex,rdkit_MinEStateIndex,rdkit_MaxAbsEStateIndex,rdkit_MinAbsES tateIndex,rdkit_qed,rdkit_MolWt,rdkit_HeavyAtomMolWt,rdkit_ExactMolWt,rdkit_NumValenceElectrons,r dkit_NumRadicalElectrons,rdkit_MaxPartialCharge,rdkit_MinPartialCharge,rdkit_MaxAbsPartialCharge,rdkit t_MinAbsPartialCharge,rdkit_FpDensityMorgan1,rdkit_FpDensityMorgan2,rdkit_FpDensityMorgan3,rdkit_B CUT2D_MWHI,rdkit_BCUT2D_MWLOW,rdkit_BCUT2D_CHGHI,rdkit_BCUT2D_CHGLO,rdkit_BCUT2 D LOGPHI,rdkit BCUT2D LOGPLOW,rdkit BCUT2D MRHI,rdkit BCUT2D MRLOW,rdkit BalabanJ,r dkit_BertzCT,rdkit_Chi0,rdkit_Chi0v,rdkit_Chi1,rdkit_Chi1n,rdkit_Chi1v,rdkit_Chi2n,rdkit_Chi 2v,rdkit_Chi3n,rdkit_Chi3v,rdkit_Chi4n,rdkit_Chi4v,rdkit_HallKierAlpha,rdkit_Ipc,rdkit_Kappa1,rdkit_Kap pa2,rdkit Kappa3,rdkit LabuteASA,rdkit PEOE VSA1,rdkit PEOE VSA10,rdkit PEOE VSA11,rdkit PE OE VSA12,rdkit PEOE VSA13,rdkit PEOE VSA14,rdkit PEOE VSA2,rdkit PEOE VSA3,rdkit PEOE VSA4,rdkit_PEOE_VSA5,rdkit_PEOE_VSA6,rdkit_PEOE_VSA7,rdkit_PEOE_VSA8,rdkit_PEOE_VSA9,rd kit_SMR_VSA1,rdkit_SMR_VSA10,rdkit_SMR_VSA2,rdkit_SMR_VSA3,rdkit_SMR_VSA4,rdkit_SMR_V SA5,rdkit SMR VSA6,rdkit SMR VSA7,rdkit SMR VSA8,rdkit SMR VSA9,rdkit SlogP VSA1,rdkit Sl ogP_VSA10,rdkit_SlogP_VSA11,rdkit_SlogP_VSA12,rdkit_SlogP_VSA2,rdkit_SlogP_VSA3,rdkit_SlogP_ VSA4,rdkit_SlogP_VSA5,rdkit_SlogP_VSA6,rdkit_SlogP_VSA7,rdkit_SlogP_VSA8,rdkit_SlogP_VSA9,rdk it_TPSA,rdkit_EState_VSA1,rdkit_EState_VSA10,rdkit_EState_VSA11,rdkit_EState_VSA2,rdkit_EState_V SA3,rdkit_EState_VSA4,rdkit_EState_VSA5,rdkit_EState_VSA6,rdkit_EState_VSA7,rdkit_EState_VSA8,rd kit_EState_VSA9,rdkit_VSA_EState1,rdkit_VSA_EState10,rdkit_VSA_EState2,rdkit_VSA_EState3,rdkit_V SA_EState4,rdkit_VSA_EState5,rdkit_VSA_EState6,rdkit_VSA_EState7,rdkit_VSA_EState8,r tate9,rdkit_FractionCSP3,rdkit_HeavyAtomCount,rdkit_NHOHCount,rdkit_NOCount,rdkit_NumAliphaticCa rbocycles,rdkit_NumAliphaticHeterocycles,rdkit_NumAliphaticRings,rdkit_NumAromaticCarbocycles,rdkit_ NumAromaticHeterocycles,rdkit_NumAromaticRings,rdkit_NumHAcceptors,rdkit_NumHDonors,rdkit_Num Heteroatoms,rdkit_NumRotatableBonds,rdkit_NumSaturatedCarbocycles,rdkit_NumSaturatedHeterocycles,rd kit_NumSaturatedRings,rdkit_RingCount,rdkit_MolLogP,rdkit_MolMR,rdkit_fr_Al_COO,rdkit_fr_Al_OH,rd kit fr Al_OH_noTert,rdkit_fr ArN,rdkit_fr Ar_COO,rdkit_fr Ar_N,rdkit_fr Ar_NH,rdkit_fr Ar_OH,rdkit_

fr_COO,rdkit_fr_COO2,rdkit_fr_C_O,rdkit_fr_C_O_noCOO,rdkit_fr_C_S,rdkit_fr_HOCCN,rdkit_fr_Imine,r dkit fr NH0,rdkit fr NH1,rdkit fr NH2,rdkit fr N O,rdkit fr Ndealkylation1,rdkit fr Ndealkylation2,rdki t_fr_Nhpyrrole,rdkit_fr_SH,rdkit_fr_aldehyde,rdkit_fr_alkyl_carbamate,rdkit_fr_alkyl_halide,rdkit_fr_allylic _oxid,rdkit_fr_amide,rdkit_fr_amidine,rdkit_fr_aniline,rdkit_fr_aryl_methyl,rdkit_fr_azide,rdkit_fr_azo,rdkit _fr_barbitur,rdkit_fr_benzene,rdkit_fr_benzodiazepine,rdkit_fr_bicyclic,rdkit_fr_diazo,rdkit_fr_dihydropyridi ne,rdkit_fr_epoxide,rdkit_fr_ester,rdkit_fr_ether,rdkit_fr_furan,rdkit_fr_guanido,rdkit_fr_halogen,rdkit_fr_hd rzine,rdkit_fr_hdrzone,rdkit_fr_imidazole,rdkit_fr_imide,rdkit_fr_isocyan,rdkit_fr_isothiocyan,rdkit_fr_keton e,rdkit_fr_ketone_Topliss,rdkit_fr_lactam,rdkit_fr_lactone,rdkit_fr_methoxy,rdkit_fr_morpholine,rdkit_fr_nit rile,rdkit_fr_nitro,rdkit_fr_nitro_arom,rdkit_fr_nitro_arom_nonortho,rdkit_fr_nitroso,rdkit_fr_oxazole,rdkit_f r_oxime,rdkit_fr_para_hydroxylation,rdkit_fr_phenol,rdkit_fr_phenol_noOrthoHbond,rdkit_fr_phos_acid,rdki t fr phos ester, rdkit fr piperdine, rdkit fr piperzine, rdkit fr priamide, rdkit fr prisulfonamd, rdkit fr pyridi ne,rdkit_fr_quatN,rdkit_fr_sulfide,rdkit_fr_sulfonamd,rdkit_fr_sulfone,rdkit_fr_term_acetylene,rdkit_fr_tetra zole,rdkit_fr_thiazole,rdkit_fr_thiocyan,rdkit_fr_thiophene,rdkit_fr_unbrch_alkane,rdkit_fr_urea,JR_BoilingP oint,JR MeltingPoint,JR CriticalTemp,JR CriticalPress,JR CriticalVolume,JR EnthalpyForm,JR GibbsEne rgy,JR_HeatCapacity,JR_EnthalpyVap,JR_EnthalpyFus,JR_LiquidVisco,JR_CrystalSolub_1,JR_CrystalSolu b_2,JR_AmorphSolub_1,JR_AmorphSolub_2,DFTMD_vol

[explanation] DFT_*: parameters calculated by DFT

rdkit_: parameters calculated by RDKit

JR_: parameters estimated by group contribution method

DFTMD_vol: molecular volume estimated by DFT-MD simulation

3: output the descriptors as list data.

Answer by GPT-4

Task 1:

The refractive index of a material is the ratio of the speed of light in a vacuum to the speed of light in that material. In general, the refractive index of molecules can be theoretically described by the Lorentz-Lorenz equation (also known as the Clausius-Mossotti relation), which is:

 $n^2 - 1 / n^2 + 2 = 4\pi/3 * N * \alpha$

where:

- n is the refractive index
- N is the number density of the molecules (number of molecules per unit volume)
- α is the polarizability of the molecules

The polarizability describes the ability of a molecule to form instantaneous dipoles, which is directly linked to the electronic structure of the molecules. Therefore, descriptors that reflect the electronic structure, size and shape of the molecules should be relevant for predicting the refractive index.

Task 2 & 3:

Based on the general principles above, the most relevant descriptors from the provided list for predicting the refractive index of aliphatic and aromatic polymers would be:

- 1. DFT_energy Reflects overall energy of the system that could correlate with polarizability.
- 2. DFT_HOMO Energy of the Highest Occupied Molecular Orbital, related to electron distribution.
- 3. DFT_LUMO Energy of the Lowest Unoccupied Molecular Orbital, also related to electron distribution.
- 4. DFT_dipoleTot Total dipole moment, directly related to polarizability.
- 5. rdkit_MolWt Molecular weight, which could relate to size of the molecule.
- 6. rdkit_NumValenceElectrons The number of valence electrons, related to the bonding and potential polarizability.
- 7. rdkit_NumAromaticRings Reflects presence of aromatic systems which could alter the refractive index.
- 8. rdkit_NumAliphaticRings Reflects presence of aliphatic rings which could alter the refractive index.
- 9. DFTMD_vol Molecular volume estimated by DFT-MD simulation, could correlate with size of the molecule.
- 10. rdkit_HeavyAtomMolWt Weight of the molecule without hydrogen atoms, could be a measure of molecular size.
- 11. rdkit_MaxAbsPartialCharge Could be linked to polarizability.
- 12. rdkit_MolLogP Describes the lipophilicity of the molecule which could correlate with polarizability.
- 13. JR_BoilingPoint Boiling point can reflect intermolecular forces and thus polarizability.
- 14. JR_EnthalpyForm Enthalpy of formation, reflects the stability of the molecule, may correlate with polarizability.

Output list:

["DFT_energy", "DFT_HOMO", "DFT_LUMO", "DFT_dipoleTot", "rdkit_MolWt", "rdkit_NumValenceElectrons", "rdkit_NumAromaticRings", "rdkit_NumAliphaticRings", "DFTMD_vol", "rdkit_HeavyAtomMolWt", "rdkit_MaxAbsPartialCharge", "rdkit_MolLogP", "JR_BoilingPoint", "JR_EnthalpyForm"]

Please note, however, that the selection of descriptors would ideally be done based on a more detailed understanding of the specific polymers under consideration and the performance of these descriptors in initial models. Additionally, a machine learning model could be employed to handle feature selection and automatically pick out the most informative descriptors.

(2) Causal relationships

Question by Author

Think step by step. Estimate the causal relationship between the following variables and output it in Mermaid syntax.

Refractive_index

DFT_energy

DFT_HOMO

DFT_LUMO

DFT_dipoleTot

rdkit_MolWt

rdkit_NumValenceElectrons

rdkit_NumAromaticRings

rdkit_NumAliphaticRings

DFTMD_vol

rdkit_HeavyAtomMolWt

rdkit_MaxAbsPartialCharge

rdkit_MolLogP

JR_BoilingPoint

JR_EnthalpyForm

Answer by GPT-4

Mermaid is a simple markdown-like syntax that lets you generate flowcharts, sequence diagrams, and other types of diagrams. To write a diagram depicting a causal relationship, we use a flowchart syntax.

However, I should note that in order to map the causal relationships between these variables accurately, I would need to have specific knowledge about the relationships themselves. In the absence of this information, I'll have to make an assumption for the purposes of demonstration.

Let's assume a hypothetical situation where:

'Refractive_index' is influenced by 'DFT_energy', 'DFT_HOMO', 'DFT_LUMO', 'DFT_dipoleTot', and 'DFTMD_vol'.

'DFT_energy', 'DFT_HOMO', 'DFT_LUMO', 'DFT_dipoleTot' and 'DFTMD_vol' are influenced by 'rdkit_MolWt' and 'rdkit_NumValenceElectrons'.

'rdkit_MolWt' and 'rdkit_NumValenceElectrons' are influenced by 'rdkit_NumAromaticRings', 'rdkit_NumAliphaticRings', 'rdkit_HeavyAtomMolWt', 'rdkit_MaxAbsPartialCharge', 'rdkit_MolLogP', 'JR_BoilingPoint', and 'JR_EnthalpyForm'.

The Mermaid syntax would be something like this:

graph LR

- JR_BoilingPoint -->|Influence| rdkit_MolWt
- JR_BoilingPoint -->|Influence| rdkit_NumValenceElectrons
- JR_EnthalpyForm -->|Influence| rdkit_MolWt
- JR_EnthalpyForm -->|Influence| rdkit_NumValenceElectrons
- rdkit_MolLogP -->|Influence| rdkit_MolWt
- rdkit_MolLogP -->|Influence| rdkit_NumValenceElectrons
- rdkit_MaxAbsPartialCharge -->|Influence| rdkit_MolWt
- rdkit_MaxAbsPartialCharge -->|Influence| rdkit_NumValenceElectrons
- rdkit_HeavyAtomMolWt -->|Influence| rdkit_MolWt
- rdkit_HeavyAtomMolWt -->|Influence| rdkit_NumValenceElectrons
- rdkit_NumAliphaticRings -->|Influence| rdkit_MolWt
- rdkit_NumAliphaticRings -->|Influence| rdkit_NumValenceElectrons
- rdkit_NumAromaticRings -->|Influence| rdkit_MolWt
- rdkit_NumAromaticRings -->|Influence| rdkit_NumValenceElectrons
- rdkit_MolWt -->|Influence| DFT_energy
- rdkit_MolWt -->|Influence| DFT_HOMO
- rdkit_MolWt -->|Influence| DFT_LUMO
- rdkit_MolWt -->|Influence| DFT_dipoleTot
- rdkit_MolWt -->|Influence| DFTMD_vol
- rdkit_NumValenceElectrons -->|Influence| DFT_energy
- rdkit_NumValenceElectrons -->|Influence| DFT_HOMO

rdkit_NumValenceElectrons -->|Influence| DFT_LUMO

- rdkit_NumValenceElectrons -->|Influence| DFT_dipoleTot
- rdkit_NumValenceElectrons -->|Influence| DFTMD_vol
- DFT_energy -->|Influence| Refractive_index
- DFT_HOMO -->|Influence| Refractive_index
- DFT_LUMO -->|Influence| Refractive_index
- DFT_dipoleTot -->|Influence| Refractive_index
- DFTMD_vol -->|Influence| Refractive_index

Table S 3 Full regression results.

Descriptor	Model	MAE	RMSE
GPT-4	Ridge	0.0229	0.0310
GPT-4	Lasso	0.0236	0.0320
All	Lasso	0.0272	0.0334
Boruta	GPR	0.0276	0.0339
Boruta	Ridge	0.0279	0.0345
All	GBR	0.0287	0.0374
All	Ridge	0.0291	0.0373
Boruta	RFR	0.0295	0.0391
All	RFR	0.0297	0.0394
Boruta	GBR	0.0304	0.0384
GPT-4	GPR	0.0304	0.0372
GPT-4	RFR	0.0311	0.0403
Boruta	Lasso	0.0314	0.0410
GPT-4	GBR	0.0328	0.0422
Random-20	RFR	0.0349	0.0448
Random-20	GBR	0.0372	0.0453
Random-10	RFR	0.0383	0.0477
Random-20	GPR	0.0391	0.0514
Random-10	GPR	0.0404	0.0512
Random-10	GBR	0.0407	0.0518
Random-10	Lasso	0.0430	0.0548
Random-20	Ridge	0.0460	0.0633

Random-20	SVR	0.0536	0.0641
Boruta	SVR	0.0546	0.0650
GPT-4	SVR	0.0553	0.0655
Random-10	SVR	0.0557	0.0662
Random-10	Ridge	0.0564	0.0844
All	SVR	0.0567	0.0669
Random-20	Lasso	0.0572	0.0820
All	GPR	0.0664	0.1310



b)

Fig. S3 Venn's diagram for the descriptor sets. a) All-Lasso and GPT-Lasso. b) GPT and GPT-Lasso.

GPT: Descriptors selected by GPT-4.

GPT-Lasso: Descriptors selected by Lasso from the GPT set.

All-Lasso: Descriptors selected by Lasso from the All set.

For GPT-Lasso and All-Lasso, all 44 polymers were used for regression.