Multi-modal cascade feature transfer for polymer property prediction

Kiichi OBUCHI[1,2]
kiichi-obuchi@nec.com

Yuta YAHAGI[1,2]
yuta-yahagi@nec.com

Kiyohiko TOYAMA[2]
k-toyama_bq@nec.com

Shukichi TANAKA[2]
shukichi.tanaka@nec.com

Kota MATSUI[3]
matsui.kota.x3@f.mail.nagoya-u-ac.jp

[1]NEC Corporation, Minato-ku, Tokyo, Japan, 211-8666
[2]National Institute of Advanced Industrial Science and Technology, Tsukuba City, Japan, 305-8568
[3]Nagoya University, Nagoya City, Japan, 466-8550

Abstract

In this paper, we put forth a multi-modal cascade model with feature transfer with the aim of adjusting the characteristics of polymer property prediction. Polymers are characterised by a composite of data in several different formats, including molecular descriptors and additive information as well as chemical structures. Our model enables more accurate prediction of physical properties for polymers by combining features extracted from the chemical structure by GCN with features such as molecular descriptors and additive information. The predictive performance of the proposed method is empirically evaluated using several polymer datasets. We report that the proposed method shows high predictive performance compared to the baseline conventional approach using a single feature.

1 Introduction

Polymers are attractive material targets with a wide range of applications. Databases for polymers are being developed mainly in the scientific community [Otsuka et al., 2011, Kim et al., 2018, Hayashi et al., 2022], and some predictive model developments for polymers have been reported [Martin and Audus, 2023, Kuenneth et al., 2022, Wu et al., 2019, Tao et al., 2021]. Such models can be used, for example, to consider the application of adaptive experimental design, an effective data-driven approach. Adaptive experimental design is a method of repeated data collection and model updating using machine learning models trained on the data, which allows target characteristics to be reached quickly.

However, the use of machine learning models in the polymer domain faces the following two challenges: (1) The cost of data collection in chemical fields is quite expensive and very often sufficient data is not accumulated, especially in the field of advanced materials. In such cases, it is difficult to pre-train high-performing machine learning models (the so-called cold start problem). (2) As polymers contain not only molecular structure but also other information (molecular weight, additional information, etc.), a multimodal model is needed that can adequately represent molecular structure and simultaneously process non-structural features such as numbers and vectors.

This paper proposes a multimodal cascade model with feature transfer as a method for developing predictive models for polymers. Our approach combines different features in a cascaded neural

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

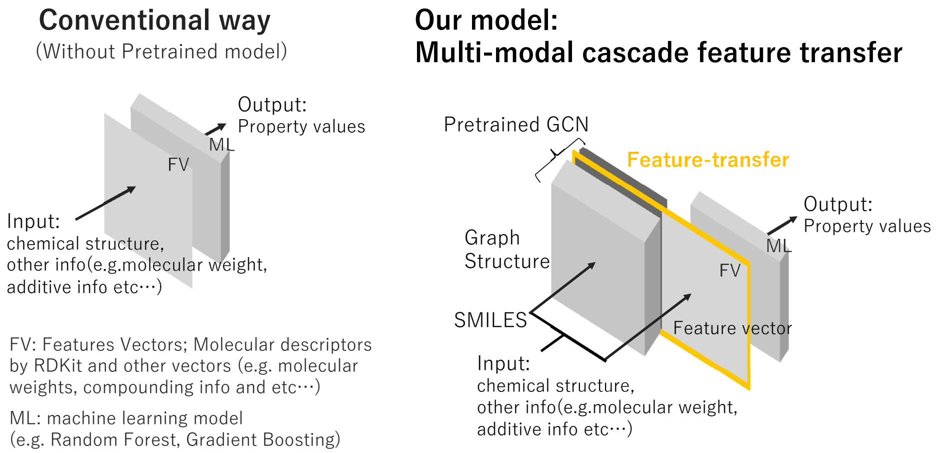

Conventional way
(Without Pretrained model)
Output:
Property values
ML
FV
Input:
chemical structure,
other info(e.g.molecular weight,
additive info etc…)
FV: Features Vectors; Molecular descriptors
by RDKit and other vectors (e.g. molecular
weights, compounding info and etc…)
ML: machine learning model
(e.g. Random Forest, Gradient Boosting)
Our model:
Multi-modal cascade feature transfer
Pretrained GCN
Feature-transfer
Output:
Property values
ML
Graph
Structure
FV
SMILES
Feature vector
Input:
chemical structure,
other info(e.g.molecular weight,
additive info etc…)

Figure 1: Conventional way (benchmark) and Our model

network structure to achieve a more effective and flexible predictive model. Furthermore, if pre-trained models can be used in the feature extractor of the cascade model, it is possible to improve the accuracy of the model in small data domains. This is expected to enable the use of big data from the public and private sectors, including past experiments, for effective modelling of private small-scale data in clandestine projects.

2 Methods

2.1 Data format and conventional approach

The chemical structures treated in this study were assumed to consist of node index: atom number, atom mass, ring and aromatic/edge attribution: bond type (single, double, triple, aromatic), ring and aromatic/adjacent matrix, and was assumed to be described in SMILES format. In addition to the chemical structure, a feature vector encoding the molecular descriptors by RDKit and information of additives encoded in an one-hot vector were used. While previous approaches have involved building a separate prediction model for each data format (left-hand side of Figure 1), this study proposes a framework in which data from all formats are processed in a unified manner using a single model.

2.2 Proposed Methods

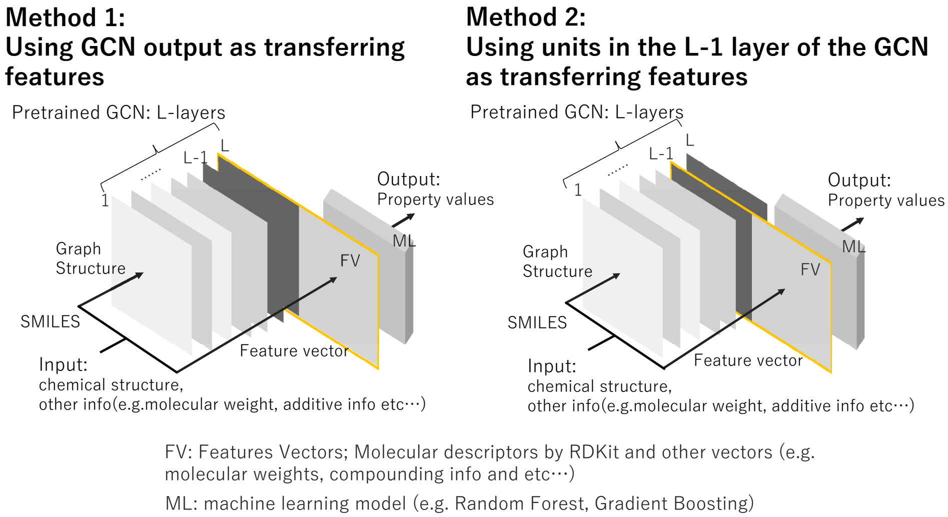
Our model is constructed in the form of a combination of two models. The first model is a feature extractor for chemical structures and in this study a graph convolutional neural network (GCN) [Schlichtkrull et al., 2018] is employed. The second model predicts physical properties for polymers by using features from GCNs together with features from molecular descriptors and additives. We represent these two models as a single model by combining them by means of a cascade structure (right-hand side of Figure 1).

Furthermore, we consider options regarding which layer of the GCN the predictive model should be coupled with. Specifically, two methods, which were attempted in this study, were formulated following Eq. (1) and Eq. (2). Method 1 combines the final layer of the GCN (L-th layer) with the predictive model, whereas Method 2 combines the layer before the final layer ($L-1$-th layer) with the predictive model:

$$\text{Method 1:} \quad \boldsymbol{h}^{(1)} = [\boldsymbol{z}_L, \boldsymbol{f}], \quad \boldsymbol{z}_L = \boldsymbol{W}_L \boldsymbol{z}_{L-1} + \boldsymbol{b}_L, \tag{1}$$

$$\text{Method 2:} \quad \boldsymbol{h}^{(2)} = [\boldsymbol{u}_{L-1}, \boldsymbol{f}], \quad \boldsymbol{u}_{L-1} = \boldsymbol{W}_{L-1} \boldsymbol{z}_{L-2} + \boldsymbol{b}_{L-1}, \tag{2}$$

where, for Method 1, $\boldsymbol{z}_L$ are feature vectors from GCN of L-th layer, $\boldsymbol{f}$ are feature vectors such as molecular descriptors by RDKit and other vectors (e.g. molecular weights, compounding info and etc.). Furthermore, $\boldsymbol{W}_L$ and $\boldsymbol{b}_L$ are the parameters of L-th layer of GCN and $[\boldsymbol{a}, \boldsymbol{b}]$ is the symbol for



Method 1:
Using GCN output as transferring features

Method 2:
Using units in the L-1 layer of the GCN as transferring features

FV: Features Vectors; Molecular descriptors by RDKit and other vectors (e.g. molecular weights, compounding info and etc···)

ML: machine learning model (e.g. Random Forest, Gradient Boosting)

Figure 2: Our model of Method 1 and Method 2

the concatenation of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. The basic symbol definition is the same in Method 2, but $\boldsymbol{u}_{L-1}$ represents the units of $L-1$-th layer of the GCN (an affine transformation of the output from the $L-2$ layer using the parameters of the $L-1$ layer). Considering such different feature transfer scenarios is an option to the empirical reports that predictive accuracy can vary for each downstream task, depending on which layer of the pre-trained model are used as transfered features [Huang et al., 2022]. Finally, we use the new feature vector $\boldsymbol{h}^{(i)}$ $i = 1, 2$ as the input of our prediction model. The two methods were illustrated in Figure 2.

2.3 Data Preparation and Cleansing

In this study, as a proof of concept of polymer case, we prepared datasets by following steps.

Polymer datasets were prepared by filtering in PoLyInfo [Otsuka et al., 2011] regarding neat resin and compound datasets of simple structure, which had Tg property. Then we operated data cleansing compounding datasets because additives information was not tidy. For example, several information were contaminated in the one cells, such as additives name, values, unit. Sometimes these ways of term separation were not uniformed. Additionally, several additives made more complicated. Under these complicated situation, rule-based cleansing didn't work well. Thus, we tried using LLM (Claude 3.5 Sonnet [Anthropic]) in order to curate these datasets interactively. It resulted in that 7503 neat resin datasets(NR datasets) and 453 compounding datasets(Comp datasets) were obtained.

NR datasets were split into train of source data(80%: 6003 datasets), val of source data(10%: 750 datasets) and target data(10%: 750 datasets). These train of source data and val of source data of NR datasets were used as building pretrained model, and target data of NR datasets and Comp datasets were split into train and val to verify our model.

2.4 Evaluation

We evaluated our model by using two target datasets; NR datasets of target and Comp datasets. As a benchmark, descriptors model using RDKit were built up. For prediction model, Gradient Boosting and Random Forest were used in both Method 1 and 2. For target data division, we divided each target datasets into the appropriate ratio of training and verification. The average R2 value of each result obtained by changing the random seed 100 times was used as an indicator to avoid large variations in the results due to the division method.

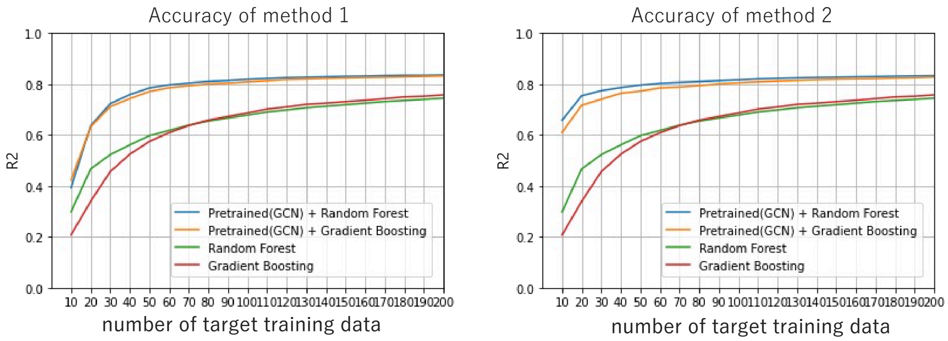


Figure 3: Results of transferred to target Neat resin datasets

3 Results and Discussion

3.1 pretrained model

The GCN was trained with GPU using approximately 6,000 source data points, with 750 data points utilized for validation purposes. the validation loss reached a plateau at approximately 500 epochs, early stopping was employed to terminate the training process.

3.2 transferred to target data of Neat Resin datasets

The results of the target NR data are presented in Fig. 3. This figure depicts the outcomes of the model validation using the target data of the NR datasets. As illustrated in Fig. 3(a), in the case of Method 1, the mean of val R2 score surpassed 0.7 for both Random Forest and Gradient Boosting with a mere 30 training data points. Moreover, the value reached approximately 0.75 with 40 training data points and approximately 0.8 with 60 training data points. In contrast, the average R2 scores of benchmark were approximately 0.45 (Gradient Boosting) and 0.52 (Random Forest) with 30 training data points. These scores exceeded 0.6 with 60 training data points and 110 training data points were required to exceed 0.7. Even 200 data points were trained, the value remained below 0.8. These results represented that mere 27% training data points were necessary, comparing our model to the benchmark based on when the accuracy reaches 0.7. As illustrated in Fig. 3(b), in the case of Method 2, the mean R2 score of val surpassed 0.6 with a mere 10 training data points and reached 0.7 with 20 training data points for both Random Forest and Gradient Boosting. The result for Random Forest was 0.8 with 50 training data points. As with Method 1, if we consider an accuracy of greater than 0.7 to be the benchmark, it can be stated that the accuracy is superior to that achieved with 110 training data points and only 20 training data points, representing 18% of the training data points. These results demonstrate that feature transfer is a more effective approach than Method 1. This is because Method 2 transfers a greater number of multidimensional vectors, which is believed to be the reason why a model with good expressive power can be constructed.

3.3 transferred to target data of Compounding datasets

The results of Method 1, both Random Forest and Gradient Boosting demonstrated a negligible impact of transition learning. Yet, Method 2 could showed the advantage of our model compared to benchmark. Specifically, the mean R2 score of the validation set in was approximately 0.07 higher than that of our model for both Random Forest and Gradient Boosting in a small data area comprising less than 50 training data points. In the case of our model, a comparison of the number of training data sets in which the average R2 score of val exceeded 0.7 revealed that the benchmark had 110 data sets, while our model had 60 data sets, representing an equivalent level of accuracy with approximately 55% of the number of data sets. These results were illustrated in Fig.4(Appendix)

4 Conclusion

This paper presents a multi-modal cascade model with feature transfer as a methodology for adaptive experimental design, with the objective of modifying the characteristics of polymer property prediction. In several trials using experimental polymer datasets as a proof of concept, it was demonstrated that the proposed methodology exhibited superior accuracy compared to the benchmark. In the best pattern, even with a mere 20 training data points, representing a mere 18% of the total training dataset, in comparison to the benchmark, which necessitates a considerably larger number of training data points, specifically 110. Subsequent research will see the model expanded to encompass the intricacies of polymer composition and the nuances of their physical properties.

Acknowledgments and Disclosure of Funding

PoLyInfo datasets was provided from NIMS (National Institute of Material Science) by following contract with them.

This work was supported by Grants-in-Aid from the Japan Society for the Promotion of Science(JSPS) for Scientific Research(KAKENHI grant nos. 20K19871 and 24K20836 to K.M.).

References

Anthropic. Claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`.

Yoshihiro Hayashi, Junichiro Shiomi, Junko Morikawa, and Ryo Yoshida. Radonpy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Computational Materials*, 8(1):222, 2022.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *International conference on machine learning*, 2022.

Chiho Kim, Anand Chandrasekaran, Tran Doan Huan, Deya Das, and Rampi Ramprasad. Polymer genome: a data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C*, 122(31):17575–17585, 2018.

Christopher Kuenneth, Jessica Lalonde, Babetta L Marrone, Carl N Iverson, Rampi Ramprasad, and Ghanshyam Pilania. Bioplastic design using multitask deep neural networks. *Communications Materials*, 3(1):96, 2022.

Tyler B Martin and Debra J Audus. Emerging trends in machine learning: a polymer perspective. *ACS Polymers Au*, 3(3):239–258, 2023.

Shingo Otsuka, Isao Kuwajima, Junko Hosoya, Yibin Xu, and Masayoshi Yamazaki. Polyinfo: Polymer database for polymeric materials design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, pages 22–29. IEEE, 2011.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 2018.

Lei Tao, Vikas Varshney, and Ying Li. Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. *Journal of Chemical Information and Modeling*, 61(11):5395–5413, 2021.

Stephen Wu, Yukiko Kondo, Masa-aki Kakimoto, Bin Yang, Hironao Yamada, Isao Kuwajima, Guillaume Lambard, Kenta Hongo, Yibin Xu, Junichiro Shiomi, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *Npj Computational Materials*, 5(1):66, 2019.

A Appendix / supplemental material

A.1 results of transferred to target data of Compounding datasets

The results for the target Comp data were illustrated in Fig.4. These related section 3.3.

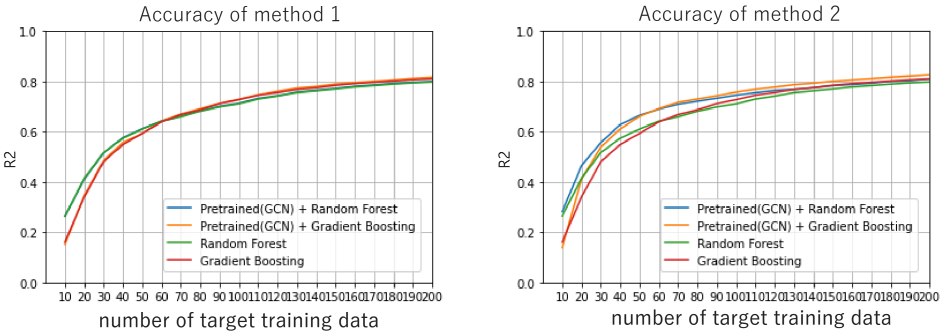


Figure 4: Results of transferred to Compounding datasets