MRS Advances © 2018 Materials Research Society
DOI: 10.1557/adv.2018.454

Number Density Descriptor on Extended-Connectivity Fingerprints Combined with Machine Learning Approaches for Predicting Polymer Properties

Takuya Minami[1], Yoshishige Okuno[1]

[1]*Research Association of High-Throughput Design and Development for Advanced Functional Materials, Ibaraki, Japan*

Abstract

We developed a new type of polymer descriptor based on Extended Connectivity Fingerprints. The number densities, that are substructure numbers divided by the number of atoms in a polymer model, were employed. We found that this approach is superior in accurately predicting linear polymer properties, compared to the conventional approach, where just the substructure numbers are used as descriptors. In addition, dimension reduction and multiple replication of repeat unit were found to improve prediction accuracy. As a result, the novel descriptor based on the Extended Connectivity Fingerprints with machine learning approaches was found to achieve accurate prediction of the refractive indices of linear polymers, which is comparable to that by ab initio density functional theory. Although process-dependent properties such as mechanical properties were difficult to predict, the present approach was found to be applicable to prediction of substructure-dependent properties, for example, optical properties, thermal stabilities.

INTRODUCTION:

Application of informatics or machine learning to material science has attracted researchers, since it accelerates the development of novel functional materials [1][2]. This is a data-driven approach, which predicts structures and/or properties of unknown materials by learning correlations between descriptors and properties of already-known materials. In recent years, researchers have succeeded in applying materials informatics to several functional materials such as thermoelectric materials [3], molecular organic light-emitting diodes [4], and low-thermal-conductivity compounds [5]. In addition, process-structure-property (PSP) linkages [6], Integrated Computational Materials Engineering (ICME) [7], as well as catalyst informatics have been reported recently [8].

One of the key for applying informatics to material science is selecting appropriate descriptors. To make computers understand a material, we need to convert features of materials into descriptors, that are computer-friendly data such as digital vectors. Since prediction accuracy depends on the quality of descriptors, the development of descriptor is very important.

In case of functional polymers, Mannodi-Kanakkithodi et al., have reported a pioneering work on designing polymer dielectrics [9]. In their approach, a polymer is represented by the chain of blocks composed of several atoms. The descriptors are the numbers of single blocks or doubly or triply linked blocks, sum of those are standardized as the total number of blocks composing the repeat unit. Their approach showed good correlation between predicted and measured values in polymer band gap and dielectric constant. However, this descriptor is difficult to be applied to arbitrary polymer systems, because the block patterns such as -CH_2- need to be defined in advance. This indicates that one of the challenges in this field is how to automatically extract descriptors from the chemical structure of polymer. Although there is automatic descriptor generation method using CORAL framework [10], the weight of the generated descriptor is difficult to interpret because it is computationally generated by using Monte Carlo simulation.

Significant research on the representations of chemical structure has been developed in the fields of chemoinformatics or cheminformatics. For example, Extended-Connectivity Fingerprints (ECFPs) [11] do not require pre-defined molecular substructures, because the descriptors are automatically generated from chemical structure. Although the ECFPs are widely used for small molecules for its versatility, their application to linear polymers has not been reported yet.

Therefore, we challenged the application of the ECFPs to linear polymers. We found the number-density ECFPs, which are the new type of descriptors interpreted as the constituent ratio of substructure, are applicable to the prediction of polymer refractive indices, in contrast to the conventional ECFPs. In addition, the prediction accuracy by our approach became comparable to that by accurate ab initio density functional theory (DFT) by performing dimensional reduction using least absolute shrinkage and selection operator (LASSO) regression [12], and by employing multiple repeat units in the calculation of descriptors. Furthermore, the limit of application of our approach is discussed based on the comprehensive study on optical properties, thermal stabilities, and mechanical properties.

THEORY:

A chemical structure was represented by the simplified molecular-input line-entry system (SMILES) [13]. To consider the substructures with respect to the linkage between repeat units, multiple repeat units were employed in the calculation of descriptors. For example, -$[CH_2O]_3$- is represented as “COCOCO” by SMILES. We obtained ECFPs [11] by counting the number of substructures from SMILES by using RDKit [14], where the maximum diameter of the circular neighborhood for each atom and the length of descriptor vector are set to be 4 and 1024, respectively. This type of ECFPs is defined as ECFP4 according to the diameter. For convenience, we call these conventional ECFP4 as the *number-ECFP4*. On the other hand, the *number-density-ECFP4* were obtained from *number-ECFP4* divided by the total number of atoms in a polymer model. The number density is defined as,

$$r_i = N_i / N, \qquad (1)$$

where N_i is the value of i-th dimension in number-ECFP4 and N is the total number of atoms in a polymer model.

Most of variables composing ECFP4 are often unnecessary for predicting polymer properties. We therefore performed the LASSO regression for dimension reduction as

follows. By using training dataset of cross-validation process, (i) hyperparameters of LASSO regression was tuned, (ii) coefficients of the regression model were obtained, and (iii) we removed variables whose coefficients are zero. The final number of descriptors was automatically decided by this procedure. In the present case, we obtained descriptors with tens of dimensions, though the dimension varies for target physical properties. After the dimension reduction, the Gaussian process regression (GPR) [15] was used to predict polymer properties. The LASSO regression and GPR were performed by using the scikit-learn [16]. We used the leave-one-out cross validation, where hyperparameter tuning and dimension reduction were performed for training dataset, and prediction accuracies were evaluated by the determination coefficient (R^2) and root mean squared error (RMSE) in test dataset. Although there are several definitions of R^2, the following definition was employed in this study,

$$R^2 \equiv 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}, \qquad (2)$$

where, $\bar{y}$ is the average of measured values, and y_i and f_i are the i-th measured and predicted values, respectively.

RESULTS AND DISCUSSIONS:

As the first case, we show the result of refractive indices of linear polymers. The 51 data were collected from the study by Maekawa, et al. [17] and Polyinfo [18]. Table I summarizes the computational conditions and results. To clarify the effectiveness of number-density-ECFP4, multiple repeat units, and dimension reduction, five examinations (a) – (e) were performed. In addition, the prediction accuracies of our machine learning approach were compared with those of the ab-initio density functional theory (DFT) conducted by Maekawa et al.

The small R^2 value (-0.063) in the column (a) of Table I indicates that the conventional number-ECFP4 failed to predict the refractive indices of linear polymers. Due to the definition (Eq 2), R^2 could takes negative value, when the numerator is larger than the denominator; herein, the negative value of R^2 means that the prediction accuracy is bad. In contrast, our number-density ECFP4 highly improved the prediction accuracy, as evidenced by the high R^2 (0.890) value and small RMSE (0.027). These results clearly show that the number-density-ECFP4 is superior to the conventional number-ECFP4 for predicting the refractive indices of linear polymers.

Here, we discuss the difference between the number- and number-density descriptors. For simplicity, we consider the descriptors for C atoms in two kinds of repeat-unit models represented as -$[CH_2CH_2O]_O$- and -$[CH_2CH_2OCH_2CH_2O]_O$-. It is noted that these two polymers are identical to one another because those chain lengths are infinite. On one hand, the conventional number-descriptor incorrectly identifies them as the different polymers from each other, since the numbers of C atoms of -$[CH_2CH_2O]$- and -$[CH_2CH_2OCH_2CH_2O]$- are 2 and 4, respectively. On the other hand, the number-density-descriptor can correctly identify them as the same polymer, since the number density of C atoms are 2/7 for both cases. This example shows that the number-density descriptor is appropriate for learning structural features on arbitrary linear polymers composed of different sized repeat units each other.

In addition to the introduction of the number-density descriptor, the prediction accuracies were improved by applying dimension reduction (c), by employing 10 repeat units to generate descriptor (d), and by combining both of them (e). The difference

between results (b) and (d) indicates that the linkage of polymer repeat unit give non-negligible effect on polymer properties, because, as mentioned in the THEORY section, multiple repeat units include the substructure of linkage. Interestingly, our machine learning approach (e) achieved similar accuracy with the DFT calculation, as shown in Table I and in the scatter plot of refractive indices in Figure 1.

Table I. Prediction conditions and accuracies of refractive indices of linear polymers

	(a)	(b)	(c)	(d)	(e)	DFT[17]
			Computational conditions			
ECFP4	Number (Original)	Number density	Number Density	Number Density	Number Density	N/A
Numbers of repeat units	1	1	1	10	10	N/A
Dimension reduction	No	No	Yes	No	Yes	N/A
			Computational results			
R^2	-0.063	0.890	0.900	0.932	0.950	0.918
RMSE	0.084	0.027	0.026	0.021	0.017	0.018

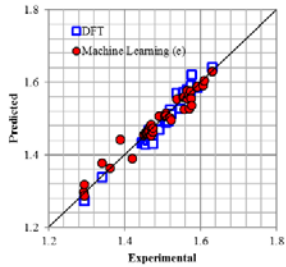

Figure 1. Scatter plot of refractive indices computed by DFT [Reprinted with permission from S. Maekawa, K. Moorthi, J. Phys. Chem. B, 120, 2507 (2016). Copyright 2016 American Chemical Society.], and by machine learning at condition (e) in Table I.

Table II shows the summary of the prediction accuracies of the optical properties (band gap, dielectric constant, and refractive index), the thermal stabilities (glass transition temperatures and linear expansion coefficients), and the mechanical properties (tensile strength at break, elongation at break, and tensile modulus). All of these calculations were carried out by the condition (e) noted in Table I. As shown in Table II, the optical properties and thermal stabilities were predictable by our approach. Because our approach effectively includes the information on polymer substructures known to depend on these properties.

However our approach was unable to predict the mechanical properties (Table II). This is because mechanical properties depend not only on substructures but also on manufacturing process of polymers [20]. In order to improve prediction accuracies of mechanical properties, an expansive informatics framework which includes not only the information concerning chemical structure but also the other information such as manufacturing process, phase, and higher-order structure will be required.

Table II. Prediction accuracies of linear polymers computed at the condition (e) noted in Table I.

	Number of data	R^2	RMSE
Optical properties			
Band gap [eV]	$284^{\$}$[9]	0.84	0.48
Dielectric constant [-]	$284^{\$}$ [9]	0.76	0.47
Refractive index [-]	51^{I}[17][18]	0.95	0.02
Thermal stabilities			
Glass transition temperature [K]	417^{I}[18]	0.84	31.1
Linear expansion coefficient [10^{-5} /K]	54^{I}[19]	0.64	1.1
Mechanical properties			
Tensile strength at break [GPa]	175^{I}[18]	-0.04	0.21
Elongation at break [%]	168^{I} [18]	0.02	196.3
Tensile modulus [GPa]	147^{I} [18]	0.03	12.7

$^{\$}$ supervised data were obtained by calculation. $^{\text{I}}$ supervised data were obtained by experiment.

CONCLUSIONS:

The conventional ECFPs were found not to be applicable to the prediction of linear polymer properties. This problem was overcome by creating the novel number-density ECFPs. We believe that structure normalization such as number density is a key concept for employing promising descriptors developed in chemoinformatics to linear polymers. For example, not only the ECFPs, but also topological torsion [21], atom-pair [22], or convolutional networks on graphs [23] should become applicable to polymers.

Furthermore, the combinations of number-density-ECFPs with (1) multiple repeat units and (2) dimension reduction with the LASSO regression was found to improve prediction accuracy of the refractive indices. As a result, our machine learning approach achieved similar prediction accuracy to the ab-initio DFT calculation.

Our machine learning approach was predictable well the optical and the thermal stabilities which are supposed to depend on substructures composing polymers. On the other hand, the mechanical properties were difficult to predict, because they depend not only on substructures but also the other factors, for example, the manufacturing process. This problem should be overcome in the future research to realize an effective design of functional polymers.

The approach shown in this study could be applicable to various polymers if the structural formula is known. In addition, this approach can be combined with other descriptors, e.g., curing time, curing temperature, or grain size. This is the advantageous feature for improving multi-scaling properties. Our results would stimulate the research on data-driven material design with artificial intelligence and machine learning on functional polymer materials.

ACKNOWLEDGEMENTS:

This work was supported by a grant from the New Energy and Industrial Technology Development Organization of Japan (P16010). We also show our appreciation to Dr.

Masaaki Kawata in the National Institute of Advanced Industrial Science and Technology for valuable discussions.

REFERENCES:

[1] K. Rajan. Materials Today, 8, 38 (2005).
[2] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim. npj comput. Mat. 3, 54 (2017).
[3] M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio, and D. R. Clarke, Chem. Mat. 25, 25911 (2013).
[4] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, A. Aspuru-Guzik, Nature Materials, 15, 1120 (2016).
[5] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, Phys. Rev. Lett., 115, 205901 (2015)
[6] S. R. Kalidindi, M. D. Graef., Annu. Rev. Mater. Res., 45, 171 (2015).
[7] J. H. Panchal, S. R. Kalidindi, D. L. McDowell, Computer-Aided Design, 45, 4 (2013).
[8] A. Yada, K. Nagata, Y. Ando, T. Matsumura, S. Ichinoseki, K. Sato, Chem. Phys. Lett., 47, 284 (2018).
[9] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, R. Ramprasad, Sci. Rep. 6:20952 (2016).
[10] P. R. Duchowicz, S. E. Fioressi, D. E. Bacelo, L. M. Saavedra, A. P. Toropova, A. A. Toropov, Chemometrics and Intelligent Laboratory Systems, 140, 86 (2015).
[11] D. Rogers, M. Hahn, J. Chem. Inf. Model. 50, 742 (2010).
[12] R. Tibshirani, J. R. Statist. Soc. B 73, 273 (2011).
[13] D. Weininger, J. Chem. Inf. Comput. Sci., 28, 31 (1988).
[14] RDKit: Open-Source Cheminformatics. Available at http://rdkit.org (accessed 15 April 2017).
[15] C. E. Rasmussen and C. K. I.Williams, *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA, 2006).
[16] Scikit-learn: Machine Learning in Python. Available at http://scikit-learn.org. (accessed 1 Oct 2017).
[17] S. Maekawa, K. Moorthi, J. Phys. Chem. B, 120, 2507 (2016).
[18] Polyinfo. Available at http://polymer.nims.go.jp (accessed 30 Oct 2017).
[19] N. Kinjo, M. Ogata, S. Numata, thermoset resin, 8, 22 (1987).
[20] H. E. H. Meijer, L. E. Govaert, Prog. Polym. Sci. 30, 915 (2005).
[21] R. Nilakantan, N. Bauman, J. Dixon, R. Venkataraghavan, J. Chem. Inf. Comput. Sci., 27, 82 (1987).
[22] R.E. Carhart, D.H. Smith, R. Venkataraghavan, J. Chem. Inf. Comput. Sci., 25, 64 (1985).
[23] D. Duvenaudy, D. Maclauriny, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *Advances in Neural Information Processing Systems,* p2215 (2015).