

# Doctoral Dissertation

## Data-science for Estimating Various Properties of Polymers Based on Monomer Unit Structure Information

Hitoshi Yamano

February, 2021

Computational Systems Biology Laboratory  
Division of Information Science  
Graduate School of Science and Technology  
Nara Institute of Science and Technology

A Doctoral Dissertation  
submitted to Graduate of Science and Technology,  
Nara Institute of Science and Technology  
In partial fulfillment of the requirements for the degree of  
Doctor of ENGINEERING

Yamano Hitoshi

Thesis Committee:

Professor Shigehiko Kanaya	(Supervisor)
Associate Professor Tomoyuki Miyao	(Co-Supervisor)
Associate Professor MD.Altaf-Ul-Amin	(Co-Supervisor)
Associate Professor Naoaki Ono	(Co-Supervisor)
Assistant Professor Alex Ming Huang	(Co-Supervisor)

# Data-science for Estimating Various Properties of Polymers Based on Monomer Unit Structure Information

Hitoshi Yamano

## Abstract

Materials informatics is the approach to develop materials using combination of materials science and informatics techniques. It should be noted that high throughput experiment is also the key to materials informatics. Nowadays, in both academic and industrial fields, there are many reports which use Materials informatics in real problem to understand mechanism, predict properties or design molecules. Therefore, the main topic in chemical industry is now changing from ‘whether informatics approach become useful’ to ‘how should informatics approach be applied to real dataset’. The reason why is the lack of enough complete dataset in real difficult problems you have to solve in industry.

In this study, I will propose how to apply data-science techniques to real incomplete dataset and obtain helpful knowledge. In this thesis, I discuss how I should apply data-science approach to small and incomplete dataset describing polymer property data. Considering the dataset of polymer property includes missing value, how it should be considered is discussed. It will be also shown that unsupervised manner is useful to understand the relationships of properties.

Using the incomplete dataset, I will propose the way to predict polymer properties by data-science approach. It is simple way based on monomer unit structure information and be able to deal with various properties. I performed to evaluate the reliability prediction models and also propose judging generalization performance with data already obtained. On the basis of above consideration, I will discuss the situation of materials informatics in industrial use in the future.

## Keywords:

Materials informatics, Polymer Properties, Predicting Property, Incomplete dataset, Evaluating Predicting Model

---

\* Doctoral Dissertation, Graduate of Science and Technology, Nara Institute of Science and Technology, February, 2021

# Contents

<b>1. Introduction.....</b>	<b>7</b>
<b>1.1 Data-science in chemical industry .....</b>	<b>7</b>
<b>1.2 Current status of predicting polymer properties.....</b>	<b>8</b>
<b>1.3 Semi-empirical or computational approaches for predicting polymer property.....</b>	<b>8</b>
<b>1.4 Data-driven approaches for predicting polymer property .....</b>	<b>9</b>
<b>1.5 Current status and problems of predicting polymer properties .....</b>	<b>10</b>
<b>1.6 What is “materials informatics”? .....</b>	<b>10</b>
<b>1.7 Constructs of this paper.....</b>	<b>11</b>
<b>2. Relationships between polymer properties using data science .....</b>	<b>12</b>
<b>2.1 Introduction .....</b>	<b>12</b>
<b>2.2 Materials and Methods.....</b>	<b>14</b>
<b>2.3 Results .....</b>	<b>15</b>
<b>2.4 Discussion .....</b>	<b>19</b>
<b>3. Predicting polymer properties using data science and evaluation approach for them.....</b>	<b>22</b>
<b>3.1 Introduction .....</b>	<b>22</b>
<b>3.2 Materials and Methods.....</b>	<b>23</b>
<b>3.3 Results .....</b>	<b>28</b>
<b>3.4 Discussion .....</b>	<b>32</b>
<b>4. Outlook of materials informatics in the future.....</b>	<b>47</b>
<b>4.1 Trend of big database in chemical field .....</b>	<b>47</b>
<b>4.2 The limitation of data usage these days and importance of methodology to interpret data .....</b>	<b>47</b>
<b>5. Conclusive remarks .....</b>	<b>49</b>
Research achievement.....	50
Acknowledgement .....	51
References.....	52

# List of Figures

Figure 1: Example of polymer properties .....	8
Figure 2: Heatmap and 2D clustering overview of polymer property dataset .....	16
Figure 3: Heatmap of correlation coefficient of polymers .....	17
Figure 4: Heatmap of correlation coefficient of properties .....	18
Figure 5: Hierarchical clustering for polymers .....	20
Figure 6: Hierarchical clustering for properties .....	20
Figure 7: Example of differ of chemical structure between polymer and monomer .....	22
Figure 8: Approach for predicting polymer properties .....	25
Figure 9: Correlation coefficient of molecular descriptors .....	29
Figure 10: The summary of training results of PLSR model ( $\rho$ , $\delta$ , $T_g$ ) .....	30
Figure 11: The scatter plots of training and test dataset ( $\rho$ , $\delta$ , $T_g$ ) .....	35
Figure 12: Hierarchical clustering of training and test dataset ( $\rho$ , $\delta$ , $T_g$ ) .....	36
Figure 13: $T^2$ and $Q$ statistics of training and test dataset based on PCA .....	37
Figure 14: The summary of test results of PLSR model ( $\rho$ , $\delta$ , $T_g$ ) .....	40
Figure 15: The test result of PLSR model ( $\rho$ ) .....	41
Figure 16: The test result of PLSR model ( $\delta$ ) .....	42
Figure 17: The test result of PLSR model ( $T_g$ ) .....	43
Figure 18: Comparison of clustering result and PLSR prediction models $R^2$ .....	46

## List of Tables

Table 1: Physical properties of polymers in PolyInfo DB .....	12
Table 2: The list of properties which the dataset include .....	15
Table 3: Polymer structures which hierarchical clustered based on property information .....	21
Table 4: Descriptors used in this work .....	26
Table 5: Comparison of PLSR models of three properties .....	31
Table 6: Chemical structures of test data set .....	33
Table 7: Comparison of PLSR models of three properties .....	39
Table 8: Prediction result of test data set .....	39
Table 9: Summary of the results of PLS prediction models .....	44
Table 10: Summary of the results of PLS prediction models (sorted) .....	45

# 1. Introduction

## 1.1 Data-science in chemical industry

Data science approach is called as “Forth Paradigm” [1]. It is known effective approach along with empirical science such as experimental approach (1<sup>st</sup> paradigm), theoretical science (2<sup>nd</sup> paradigm) and computational science (3<sup>rd</sup> paradigm). It has been applied to many fields. [2-8]

This trend also occurred in chemical industry [9-14]. Soft sensor [15-16] is widely used to control status of chemical plant by constructing statistical model to predict process variables which are difficult to measure directly. It is important topics in chemical industry that how to improve the model accuracy of soft sensor and how to apply them to real systems. [17-18]

A lot of researches were reported in the field of data driven approach based on chemical structure information. Funatsu et al. reported how to design novel molecules which show high activity as drugs. The research proposes visualizing chemical space and generating chemical structures based on data-science approach. [19]

Kanaya and Eguchi et al. reported it is possible that clustering alkaloids and predict their biosynthesis pathways using graph convolutional neural networks based on chemical structure information. [20]

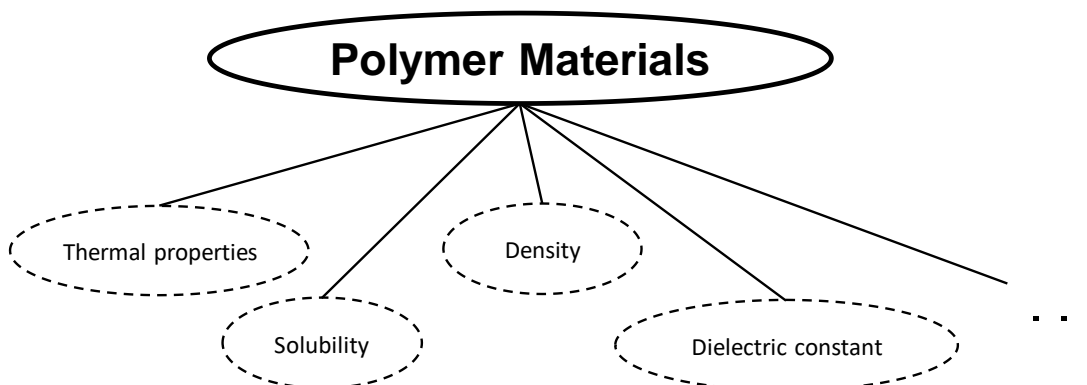
As I've mentioned so far, data-science approach has become major method in industry where enough data can be obtained. However, some fields remain difficult for applying data science because lack of perfect dataset. Polymer property is also located in one of examples. I will discuss it in the next chapter.

## 1.2 Current status of predicting polymer properties

In many chemical fields, property prediction is the most important topics to develop chemical products because reduction and optimization of experimental procedures. There are many works reported which apply data-science approach for predicting properties of small compounds [21-24]. However, predicting property of polymer materials includes specific problems which does not appear in small compounds.

Polymers are used in a wide variety of applications. It is often necessary to optimize multiple physical / chemical properties as shown in Figure1. For example: thermal properties, solubility, density, dielectric constant and so on. Polymer material development usually means to optimize these multiple different properties simultaneously.

Polymer structure is more complex than monomer structure because its property depends on not only chemical monomer design but also higher dimensional structure features and complex identity depends on its synthesis conditions such as tacticity of monomer units, distribution of molecular weight, structures of end group and so on [25-28]. Then, generally it is more difficult to predict properties of polymer structure than monomer structure.



**Figure 1: Example of polymer properties**

Polymer materials have a lot of properties and they are sometimes should be optimized simultaneously in industrial products development.

## 1.3 Semi-empirical or computational approaches for predicting polymer property

Atomic group contribution method assumes that some unique partial chemical structure (i.e., atomic group) in a compound make a certain contribution to a property and estimates property of the compound by adding them together, which is widely used approach in property prediction [29-33]. Atomic group contribution method has been also used to



estimate the properties of polymers. A lot of properties were studied by the method [34, 35], there are some empirical parameters in their models and no unified method to explain the different properties has been established [33, 36].

Predicting the properties of polymers through computational simulations have also been actively studied [37]. For example, the method of molecular dynamics (MD) [38, 39] deals with the motion of molecules and constructs force field models to simulate the behavior of polymers [40-44]. The calculations are mainly based on physical laws without considering chemical reactions. The approach is strong to solve or visualize behavior of polymer materials. However, it sometimes has some difficulty in setting parameters to explain real-world experimental results. It needs very high computational costs when the system to be handled is large.

#### **1.4 Data-driven approaches for predicting polymer property**

Otherwise, with the development of data science, the quantitative structure property relationship (QSPR) approach has been studied to predict properties of polymers based on structural information effectively. For example, glass transition temperature [45-49], pyrolysis temperature [50], refractive index [51], dielectric constant [48], and intrinsic viscosity [50, 52] have been examined based on QSPRs. There are reports that topological descriptors can be used to infer the properties [54, 55].

Ramprasad et al. proposed an approach which is based on force fields developed with machine learning methods with quantum mechanics [53]. They constructed prediction models for properties of polymers based on data including computationally formed data (bandgap [54], dielectric constant [54-57], refractive index [58], and atomization energy) or experimentally obtained data (such as glass transition temperature [59] and solubility [60, 61]). They used density functional theory (DFT) calculation and form training dataset. Prediction were based on specific fingerprint [62] including different dimensional descriptors: atomic level descriptors, QSPR descriptors and morphological descriptors. Recursive feature elimination (RFE) was used for decrease dimension of dataset. Gaussian Process Regression (GPR) was carried out to construct nonlinear prediction model to predict properties of polymers with high accuracy. The models can be used for free as “Polymer Genome” platform ([www.polymergenome.org](http://www.polymergenome.org)) [63]. You can get predicted values when you input structural information (Simplified Molecular Input Line Entry System abbreviated as SMILES) of polymer.

Oyaizu et al. [64] reported that solid polymer electrolyte material was found by machine learning approach using newly constructed database. They synthesized solid polymer electrolytes for lithium-ion battery.

Yoshida et al. [65-67] reported the Machine Learning framework called “transfer learning” is effective for property prediction of polymer materials. Pre-trained models library based on big database were “transferred” to other property which has only small dataset. They used the library comprises more than 140,000 models already constructed for various properties of small molecules, polymers, and inorganic crystalline materials. Along with the pre-trained models, they succeeded in transfer learning in different scenarios such as building models with only dozens of materials data.

### **1.5 Current status and problems of predicting polymer properties**

As mentioned above, various studies have been conducted to predict the properties of polymers using enough computational resource and/or large dataset. Otherwise, it is still difficult to predict the properties using a practical data (small amount and incomplete) effectively. From the industrial point of view, it would be useful to establish a methodology how to use or interpret the practical (small data amount and/or including missing values) dataset on the properties of polymer materials. Moreover, it is much valuable if you can develop a manner to predict desired properties applying data-science approach even if usable data is small.

### **1.6 What is “materials informatics”?**

Materials informatics is the approach to develop materials using combination of materials science and informatics techniques. It is said that using data-science approach, new values will be obtained which cannot come from materials science only. It became major after the Materials Genome Initiative project in USA [68, 69].

Especially, in the field of inorganic materials research, there are reports indicate significant success in combination of materials science and data science [72]. It should be noted that high throughput experiment is also the key to materials informatics [70-73].

Funatsu et al. proposed procedure which include constructing QSPR/QSAR models, analyze their applicability domains and generate chemical structures which have preferable property [74].

Morikawa et al. reported machine learning approach was effective to develop polymers with high thermal conductivity [67, 75, 78]. They used transfer learning and solve the problem on the lack of experimental data by using rich open dataset.

Nowadays, in both academic and industrial fields, there are many reports which use Materials Informatics in real problem to understand mechanism, predict properties or design molecules [77-81]. Therefore, the main topic in chemical industry is now changing from ‘whether informatics approach become useful’ to ‘how should informatics approach

be applied to real dataset'. The reason why is the lack of enough complete dataset in real difficult problems you have to solve in industry. In this study, I will propose how to apply data-science techniques to real incomplete dataset and obtain helpful knowledge.

## **1.7 Constructs of this paper**

In this thesis, I will discuss how I should apply data-science approach to small and incomplete dataset describing polymer property data. In chapter 2, considering the dataset of polymer property includes missing value were prepared and how it should be considered is discussed. It will be also shown that unsupervised manner is useful to understand the relationships of properties. In chapter 3, using the incomplete dataset, I will propose the way to predict polymer properties by data-science approach. It is simple way based on monomer unit structure information and be able to deal with various properties. I performed to evaluate the reliability prediction models and also propose judging generalization performance with data already obtained. The approach bases on the combination of unsupervised and supervised learning method. In chapter 4, based on above consideration, I will discuss the situation of materials informatics in industrial use in the future. Finally, I will remark my opinion concerning to feature perspective in chapter 5.

## 2. Relationships between polymer properties using data science

### 2.1 Introduction

There have been reported various types of polymer properties. For example, public database concerning PolyInfo DB (<https://polymer.nims.go.jp/>) consists of polymer properties concerning 12,913 homo-polymers, 5,537 copolymers, and 1,851 polymer blends, but it should be noted that values for the most of polymer properties are lacking because their experiments are limited. So we should compare relationships between polymer properties and between polymers taking into consideration that how missing values should be treated.

In the present study, I examined 48 polymers and 34 properties in *The Properties of Polymers* (D.W. V. Krevelen) [34] included 641 polymer species and 171 physical properties. In this section we compare the polymers and properties based on 2D-heatmap based on the pairwise correlation coefficients.

**Table 1:** Physical properties of polymers in PolyInfo DB  
(<https://polymer.nims.go.jp/>)

Group of polymer property	Polymer property
Physical properties (2)	Density, Specific volume
Optical properties (2)	Refractive index, Stress-optical coefficient
Thermal properties (14)	Crystallization kinetics, Crystallization temp., Glass transition temp.: Tg), Heat of crystallization, Heat of fusion, Thermal decomposition temp., LC phase transition temp., Linear expansion coefficient, Melting temp., Specific heat capacity: Cp, Specific heat capacity: Cv, Thermal conductivity, Thermal diffusivity, Volume expansion coefficient
Electrical	Dielectric breakdown voltage, Dielectric const.: DC, Electric

properties (5)	conductivity, Surface resistivity, Volume resistivity
Physical and chemical properties (9)	Contact angle, Gas diffusion coefficient: D, Gas permeability coefficient: P, Gas solubility coefficient: S, Hansen parameters (delta-d: dispersive component, delta-h: hydrogen bonding component, delta-p: polar component), Interfacial tension, Solubility parameter, Surface tension, Water absorption, Water vapor transmission
Dilute solution properties (10)	Diffusion coeff., Solvent / Non-solvent / Poor-solvent, Theta-solvent/theta-temp., Intrinsic viscosity: $\eta$ , Radius of gyration, Second virial coeff., Sedimentation coeff.
Rheology (2)	Dynamic viscosity, Melt viscosity
Elongation properties (7)	Dynamic tensile properties, Elongation at break, Fiber tensile elongation at break, Fiber tensile modules, Fiber tensile stress[strength] at break, Tensile modulus, Tensile stress [strength] at break, Tensile stress[strength] at yield
Shear properties (4)	Dynamic shear properties, Shear modulus, Shear stress[strength] at break, Shear stress[strength] at yield
Flexural properties (4)	Dynamic flexural properties, Flexural modulus, Flexural stress[strength] at break, Flexural stress[strength] at yield
Compressive properties (4)	Compressive modulus, Compressive stress[strength] at break, Compressive stress[strength] at yield, Dynamic compressive properties
Creep properties (12)	Compressive creep rupture strength, Compressive creep rupture time, Compressive creep strain, Flexural creep rupture strength, Flexural creep rupture time, Flexural creep strain, Tensile creep compliance, Tensile creep modulus, Tensile creep recovery, Tensile creep rupture strength, Tensile creep rupture time, Tensile creep strain
Temperature data (4)	Brittleness temp., Temperature of deflection under load, Softening temp., Vicat softening temp.
Hardness properties (2)	Rockwell hardness, Shore hardness
Fire resistance (3)	Oxygen index, UL flammability code rating, UL temp. index
Others (5)	Bulk modulus, Compressibility, G-value, PVT relation, Radiation resistance

## 2.2 Materials and Methods

*The Properties of Polymers* (D.W. V. Krevelen) [34] describes much information of general polymers. On this work, the data of polymer name, structure and their various properties were collected from the literature.

From the tables in this book, there were 641 polymer species and 171 physical properties. However, there were many overlapping including alias names, it is needed to be cleaned. After data cleansing, 48 polymers and 34 properties (including missing values) are selected. About experimental values, some properties be described several different experimental values. In such cases, the averaged values were used. Table 1 shows a list of physical properties.

I applied Ward's clustering method [82] to comprehensively understand relationships among polymer property data and those among the polymers by the property data.

**Table 2:** The list of properties which the dataset include.

No.	Symbol	unit	Property	The number of data
1	CpExpLiq	$\text{Jkg}^{-1}\text{K}^{-1}$	Heat capacity in the liquid state (298K, exp)	24
2	CpExpSolid	$\text{Jkg}^{-1}\text{K}^{-1}$	Heat capacity in the solid state (298K, exp)	30
3	CpShawLiq	$\text{Jkg}^{-1}\text{K}^{-1}$	Heat capacity in the liquid state by Shaw (298K)	30
4	CpSatoSolid	$\text{Jkg}^{-1}\text{K}^{-1}$	Heat capacity in the solid state by Satoh (298K)	32
5	delta	$\text{J}^{1/2}\text{cm}^{-3/2}$	Solubility parameter (calc)	31
6	deltaexpmax	$\text{J}^{1/2}\text{cm}^{-3/2}$	Solubility parameter (exp, max.)	23
7	deltaexpmin	$\text{J}^{1/2}\text{cm}^{-3/2}$	Solubility parameter (exp, min.)	26
8	DHm	$\text{kJ/mol}$	Molar enthalpy of fusion	30
9	DHmexp	$\text{kJ/mol}$	Molar enthalpy of fusion (exp)	21
10	DSmexp	$\text{Jmol}^{-1}\text{K}^{-1}$	Entropy of fusion (exp)	21
11	eg	$10^{-4}\text{cm}^3\text{g}^{-1}\text{K}^{-1}$	Thermal expansivity of a glass	31
12	Egcalc	$10^{-4}\text{cm}^3\text{mol}^{-1}\text{K}^{-1}$	Molar thermal expansivity of a glass	35
13	Egexp	$10^{-4}\text{cm}^3\text{mol}^{-1}\text{K}^{-1}$	Molar thermal expansivity of a glass	25
14	el	$10^{-4}\text{cm}^3\text{g}^{-1}\text{K}^{-1}$	Thermal expansivity of a liquid	33
15	Elcalc	$10^{-4}\text{cm}^3\text{mol}^{-1}\text{K}^{-1}$	Molar thermal expansivity of a liquid	35
16	Elexp	$10^{-4}\text{cm}^3\text{mol}^{-1}\text{K}^{-1}$	Molar thermal expansivity of a liquid	31
17	epsilon	-	Dielectric constant	31
18	gamma	$\text{mN/m}$	Surface tension	32
19	gammacoh	$\text{mN/m}$	Surface tension (calculated by cohesive energy density)	27
20	gammacr	$\text{mN/m}$	Critical surface tension of wetting	29
21	gammav	$\text{mN/m}$	Surface tension (calculated by parachor)	30
22	Ktheta	$\text{cm}^3\text{mol}^{1/2}\text{g}^{-3/2}$	Unperturbed viscosity coefficient (exp)	24
23	M	$\text{g/mol}$	Molar mass (molecular weight)	48
24	n	-	Index of refraction (exp)	36
25	nRGD	-	Index of refraction (calc) according to Gladstone and Dale	24
26	nRLL	-	Index of refraction (calc) according to Lorentz and Lorenz	24
27	nRV	-	Index of refraction (calc) according to Vogel	24
28	rhoa	$\text{g/cm}^3$	Density of amorphous polymer	47
29	rhoc	$\text{g/cm}^3$	Density of crystalline polymer	36
30	rhorr	$\text{g/cm}^3$	Density of rubbery amorphous polymer	41
31	Tg	K	Glass–rubber transition temperature	47
32	Tm	K	Crystalline melting point (exp)	39
33	Vr	$\text{cm}^3/\text{mol}$	Molar volume of rubbery amorphous polymer	41
34	Vw	$\text{cm}^3/\text{mol}$	Van der Waals volume	46

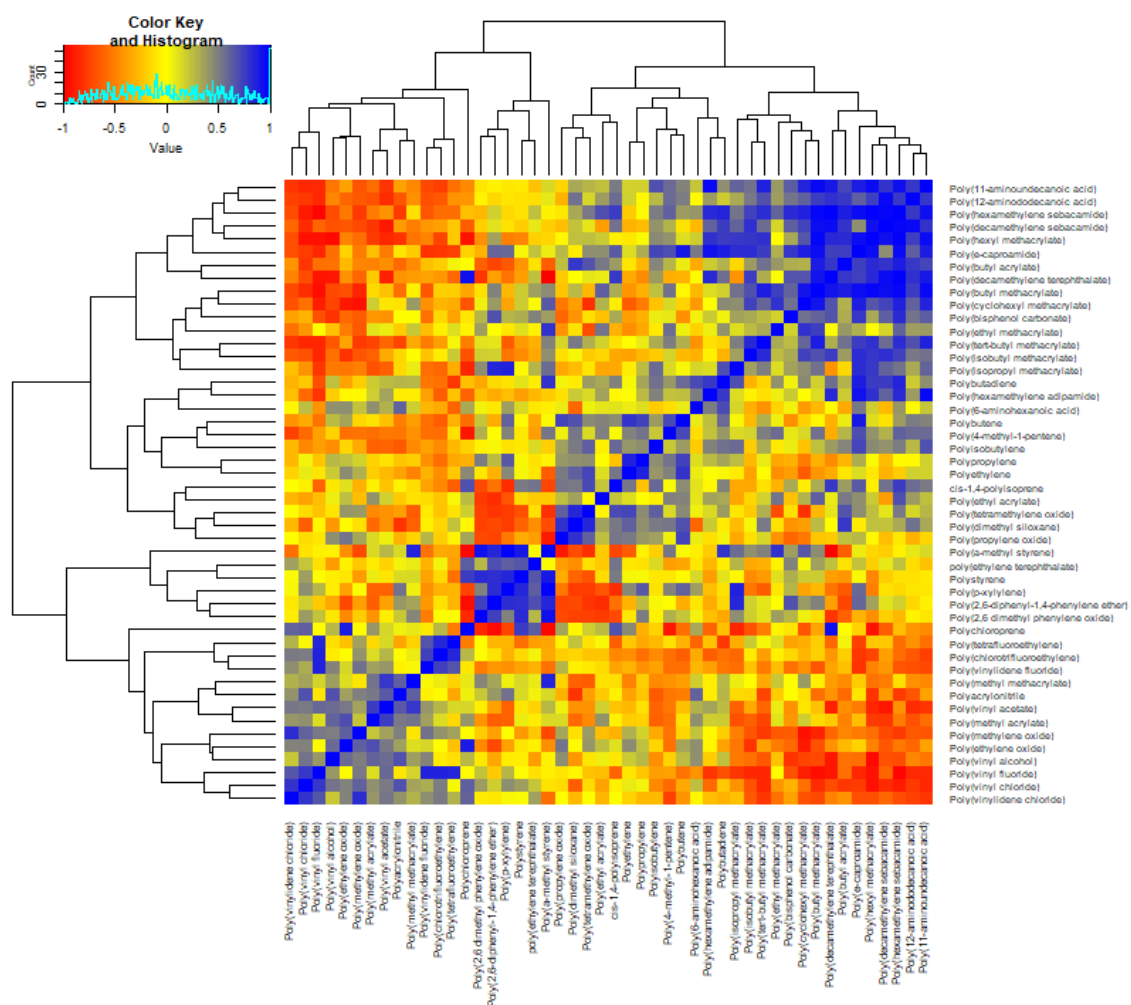
## 2.3 Results

The data matrix consists of 1632 elements, i.e., 48 x 34. The problem of the polymer data is that 554 are missing values (= 34% of the data) in the matrix. This situation occurs generally, but we cannot compare between polymer and between properties based on pairwise simple correlation methods.

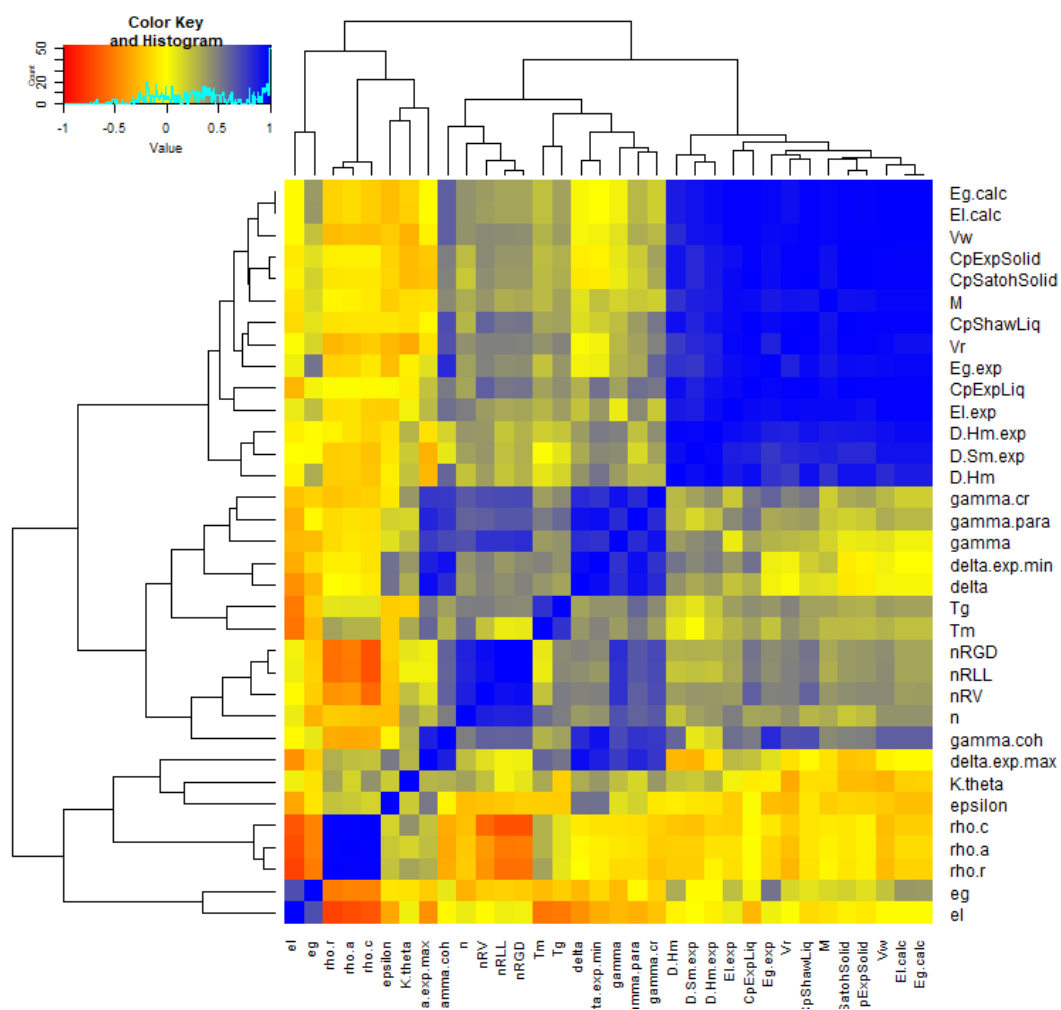
I applied Ward's clustering method [82] to comprehensively understand relationships among 34 property data of 48 polymers and those among the polymers of the property data. Hierarchical clustering can be carried out dataset with missing values like in this case.







**Figure 3: Heatmap of correlation coefficient of polymers**  
2D-clustering with heatmap for polymers



**Figure 4: Heatmap of correlation coefficient of properties**

2D-Clustering with heatmap for polymer properties

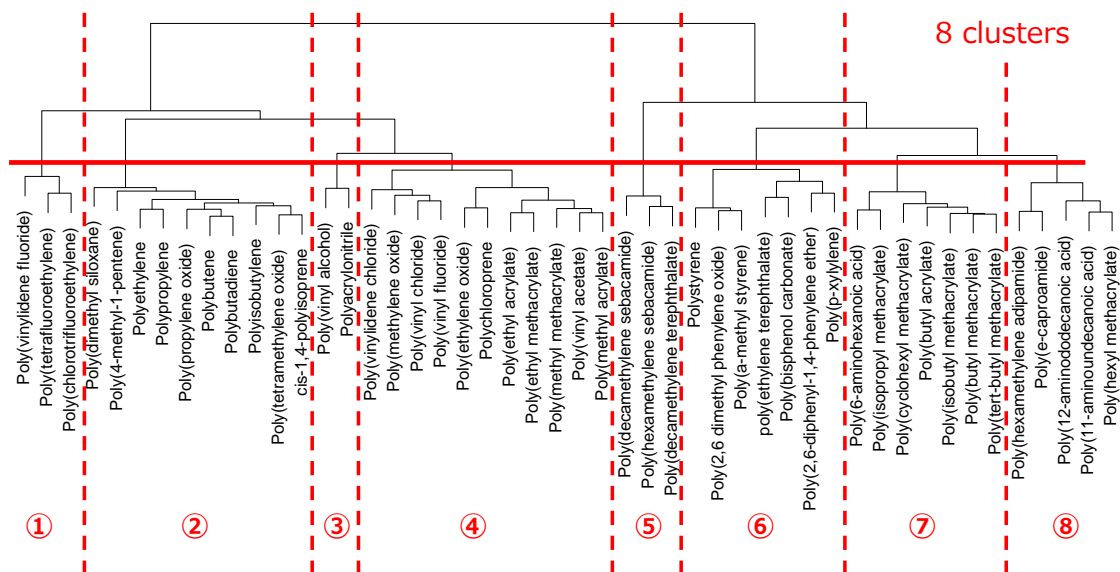
## 2.4 Discussion

### Considering relationships of polymers by hierarchical clustering

In Figure 5, closely related polymers in the 34-property dimensional space are grouped in close clusters, in contrast, those with greatly different properties are placed in different clusters. Note that this analysis is based on only property dataset without any chemical structure information. Polymers were clustered to 8 groups. In table 2, the structures of polymers of each cluster are shown. Considering this clustering result, in same cluster, structures of polymers tend to be similar. For example, in cluster no.1, there are polymers which include halogen atoms. In cluster no.2 and 3, the repeating unit of polymer is small alkyl structure. In cluster no.4, main chain is made by few carbon atoms. Cluster no.5 is made from polymers whose main chain includes more than 10 carbon atoms. The structures in cluster no.6 contain benzene ring. No.7 and 8 contains methacrylate and amide polymers. Throughout, in other clusters, similar structure polymer tends to gather same cluster.

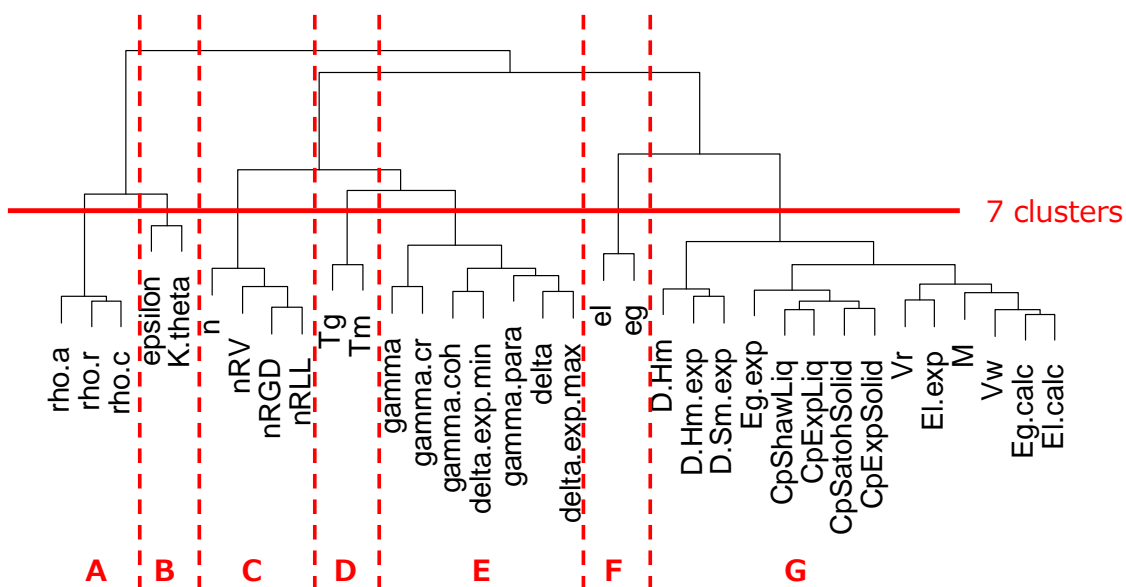
As I mentioned, this clustering result was based on the information of properties only. This approach is useful to understand the relationships of polymers using property data in data-driven method. It could be carried out for the data include missing values.

Besides, using same manner, properties of polymers were clustered into 7 groups shown in Figure 6. It helps considering and evaluating prediction models of properties. I will discuss detail on chapter 1-4.



**Figure 5: Hierarchical clustering for polymers**

Hierarchical clustering for 48 polymers based on 34 polymer properties.



**Figure 6: Hierarchical clustering for properties**

Hierarchical clustering for 34 properties based on 47 polymers.

**Table 3:** Polymer structures which hierarchical clustered based on property information

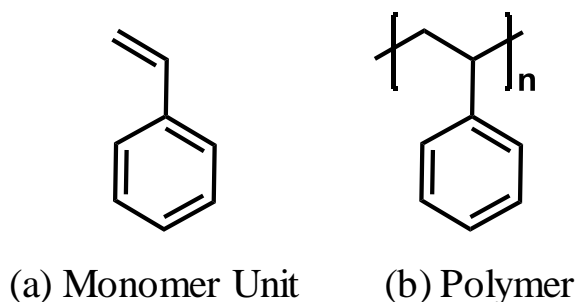
Cluster No.	Structure
1	
2	
3	
4	
5	
6	
7	
8	

### 3. Predicting polymer properties using data science and evaluation approach for them

#### 3.1 Introduction

It is difficult problem how to describe polymer structure. In the field of materials informatics, there is a lot of method to obtain molecular descriptors for polymers have been proposed but generally it is difficult to directly describe the polymer structure.

It should be needed to consider that the characteristics of the repeating structures mainly determine the character of polymers. In this section, I tried to use monomer unit structure instead of polymer structure to estimate polymer properties. Figure 7 shows a comparative example of monomer and polymer structure. There are differences between monomer units and polymer structures (such as double-bond carbons exists or not in Figure 7) and the effect of end-group unit is ignored. However, it is simple approach worth considering.



**Figure 7: Example of differ of chemical structure between polymer and monomer**

Chemical structures differs between polymer and monomer. Polymer structure is more complicated than monomer unit structure because it has repeating of the unit.

### 3.2 Materials and Methods

Figure 8 summarized the conceptual representation of estimation of targeted properties by molecular descriptors. The monomer unit structure was converted to the SMILES [83] structure formula. SMILES is structural expression notation that expresses a chemical structure by one-dimensional string information, and is often used in the field of chemoinformatics. It is difficult to express all information of polymer structure directly but it is easy in the case of monomer structure. Using SMILES information, I generated molecular descriptors by alvaDesc [84], and selected constitutional and topological descriptors consisting of 127 descriptors listed in Table 3. These descriptors were used as explanatory variables to predict polymer properties. In this study, I targeted three properties density ( $\rho$ ), dissolution parameter ( $\delta$ ), and glass transition temperature ( $T_g$ ). I performed partial least regression method for estimating the three properties by the 127 descriptors.

The PLS method has been widely used in medical imaging as well as the chemo- and bio-informatics fields because PLS models can be constructed even if there are more variables than observations. In addition, this method can be applied if multi-collinearities are hidden between the independent variables. The objective variable,  $Y$ , corresponds to the three targeted properties and the interpretive variables  $X_1, X_2, \dots, X_M$  corresponds to 127 descriptors were correlated by a linear model as Eq. (1)

$$Y = a_0 + a_1 X_1 + \dots + a_j X_j + \dots + a_M X_M \quad (1)$$

Here  $M$  represents the total number of the questions.

The PLS model is represented in *Eqs* (2) and (3).

$$\mathbf{y} = \bar{\mathbf{y}} + \sum_{k=1}^A \mathbf{t}_k q_k + \mathbf{e} = \bar{\mathbf{y}} + \mathbf{T} \cdot \mathbf{q} + \mathbf{e} \quad (2)$$

$$\mathbf{X} = \bar{\mathbf{X}} + \sum_{k=1}^A \mathbf{t}_k \mathbf{p}_k^T + \mathbf{E} = \bar{\mathbf{X}} + \mathbf{T} \cdot \mathbf{P}^T + \mathbf{E} \quad (3)$$

where  $q_k$  is the coefficient of  $y$  for the  $k^{\text{th}}$  component,  $p_k$  is the loading vector of  $X$ ,  $A$  is the number of components, and  $t_k$  is a score vector for the  $k^{\text{th}}$  component. The residual matrix and vector are represented by  $\mathbf{E}(M \times N)$  and  $\mathbf{e}(M \times 1)$ , respectively. *Eqs* (2) and (3) can be combined to create *Eq* (4).

$$\mathbf{Y} = \bar{\mathbf{y}} - \bar{\mathbf{X}}^T \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q} + \mathbf{X}^T \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q} \quad (4)$$

The number of PLS components was determined by maximizing the  $Q^2$ , which was calculated by a leave one out cross-validation for each component, as shown in Eq (5).

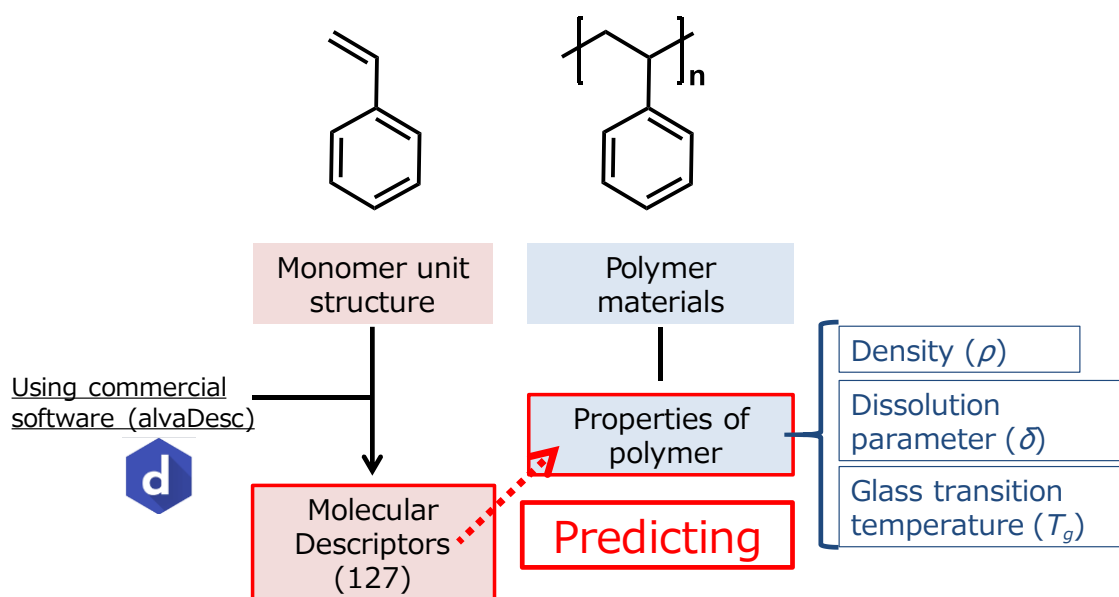
$$Q^2 = 1 - \frac{\sum_{i=1}^N (y^{(i)} - y_{cv}^{(i)})^2}{\sum_{i=1}^N (y^{(i)} - \bar{y})^2} \quad (5)$$

Here,  $y$  and  $y_{cv}^{(i)}$  are original and predicted  $y$ -values in the cross-validation for every  $i$ th individual, respectively and  $\bar{y}$  represents the average for all  $y$ -values. We determined the number of components so that  $Q^2$  value reaches the maximum. Then after determining the number of components, we also calculated the  $R^2$  for examining prediction accuracy for the PLS model.

$$R^2 = 1 - \frac{\sum_{i=1}^N (y^{(i)} - y_{all}^{(i)})^2}{\sum_{i=1}^N (y^{(i)} - \bar{y})^2} \quad (6)$$

Here,  $y_{all}^{(i)}$  represents the predicted  $y$ -value for the  $i$ th individual when the PLS model using all individuals in selecting the number of components by  $Q^2$ .





**Figure 8: Approach for predicting polymer properties**

This figure shows overview of the approach for predicting polymer properties using monomer unit structure information in this study.

**Table 4: Descriptors used in this work**

They were calculated by commercial software (alvaDesc) [84]

No.	Name	Description	Block	Sub-Block
1	MW	molecular weight	Constitutional indices	Basic descriptors
2	AMW	average molecular weight	Constitutional indices	Basic descriptors
3	Sv	sum of atomic van der Waals volumes (scaled on Carbon atom)	Constitutional indices	Basic descriptors
4	Se	sum of atomic Sanderson electronegativities (scaled on Carbon atom)	Constitutional indices	Basic descriptors
5	Sp	sum of atomic polarizabilities (scaled on Carbon atom)	Constitutional indices	Basic descriptors
6	Si	sum of first ionization potentials (scaled on Carbon atom)	Constitutional indices	Basic descriptors
7	Mv	mean atomic van der Waals volume (scaled on Carbon atom)	Constitutional indices	Basic descriptors
8	Me	mean atomic Sanderson electronegativity (scaled on Carbon atom)	Constitutional indices	Basic descriptors
9	Mp	mean atomic polarizability (scaled on Carbon atom)	Constitutional indices	Basic descriptors
10	Mi	mean first ionization potential (scaled on Carbon atom)	Constitutional indices	Basic descriptors
11	GD	graph density	Constitutional indices	Basic descriptors
12	nAT	number of atoms	Constitutional indices	Basic descriptors
13	nSK	number of non-H atoms	Constitutional indices	Basic descriptors
14	nTA	number of terminal atoms	Constitutional indices	Basic descriptors
15	nBT	number of bonds	Constitutional indices	Basic descriptors
16	nBO	number of non-H bonds	Constitutional indices	Basic descriptors
17	nBM	number of multiple bonds	Constitutional indices	Basic descriptors
18	SCBO	sum of conventional bond orders (H-depleted)	Constitutional indices	Basic descriptors
19	RBN	number of rotatable bonds	Constitutional indices	Basic descriptors
20	RBF	rotatable bond fraction	Constitutional indices	Basic descriptors
21	nDB	number of double bonds	Constitutional indices	Basic descriptors
22	nTB	number of triple bonds	Constitutional indices	Basic descriptors
23	nAB	number of aromatic bonds	Constitutional indices	Basic descriptors
24	nH	number of Hydrogen atoms	Constitutional indices	Basic descriptors
25	nC	number of Carbon atoms	Constitutional indices	Basic descriptors
26	nN	number of Nitrogen atoms	Constitutional indices	Basic descriptors
27	nO	number of Oxygen atoms	Constitutional indices	Basic descriptors
28	nP	number of Phosphorous atoms	Constitutional indices	Basic descriptors
29	nS	number of Sulfur atoms	Constitutional indices	Basic descriptors
30	nF	number of Fluorine atoms	Constitutional indices	Basic descriptors
31	nCL	number of Chlorine atoms	Constitutional indices	Basic descriptors
32	nBR	number of Bromine atoms	Constitutional indices	Basic descriptors
33	nI	number of Iodine atoms	Constitutional indices	Basic descriptors
34	nB	number of Boron atoms	Constitutional indices	Basic descriptors
35	nHM	number of heavy atoms	Constitutional indices	Basic descriptors
36	nHet	number of heteroatoms	Constitutional indices	Basic descriptors
37	nX	number of halogen atoms	Constitutional indices	Basic descriptors
38	H%	percentage of H atoms	Constitutional indices	Basic descriptors
39	C%	percentage of C atoms	Constitutional indices	Basic descriptors
40	N%	percentage of N atoms	Constitutional indices	Basic descriptors
41	O%	percentage of O atoms	Constitutional indices	Basic descriptors
42	X%	percentage of halogen atoms	Constitutional indices	Basic descriptors
43	nCsp3	number of sp <sup>3</sup> hybridized Carbon atoms	Constitutional indices	Basic descriptors
44	nCsp2	number of sp <sup>2</sup> hybridized Carbon atoms	Constitutional indices	Basic descriptors
45	nCsp	number of sp hybridized Carbon atoms	Constitutional indices	Basic descriptors
46	max_conj_path	maximum number of atoms that can be in conjugation with each other	Constitutional indices	Basic descriptors
47	nStructures	number of disconnected structures	Constitutional indices	Basic descriptors
48	totalcharge	total charge	Constitutional indices	Basic descriptors
49	ZM1	first Zagreb index	Topological indices	Vertex degree-based indices
50	ZM1V	first Zagreb index by valence vertex degrees	Topological indices	Vertex degree-based indices
51	ZM1Kup	first Zagreb index by Kupchik vertex degrees	Topological indices	Vertex degree-based indices
52	ZM1Mad	first Zagreb index by Madan vertex degrees	Topological indices	Vertex degree-based indices
53	ZM1Per	first Zagreb index by perturbation vertex degrees	Topological indices	Vertex degree-based indices
54	ZM1MulPer	first Zagreb index by multiplicative perturbation vertex degrees	Topological indices	Vertex degree-based indices
55	ZM2	second Zagreb index	Topological indices	Vertex degree-based indices
56	ZM2V	second Zagreb index by valence vertex degrees	Topological indices	Vertex degree-based indices
57	ZM2Kup	second Zagreb index by Kupchik vertex degrees	Topological indices	Vertex degree-based indices
58	ZM2Mad	second Zagreb index by Madan vertex degrees	Topological indices	Vertex degree-based indices
59	ZM2Per	second Zagreb index by perturbation vertex degrees	Topological indices	Vertex degree-based indices
60	ZM2MulPer	second Zagreb index by multiplicative perturbation vertex degrees	Topological indices	Vertex degree-based indices
61	ON0	overall modified Zagreb index of order 0	Topological indices	Vertex degree-based indices
62	ON0V	overall modified Zagreb index of order 0 by valence vertex degrees	Topological indices	Vertex degree-based indices
63	ON1	overall modified Zagreb index of order 1	Topological indices	Vertex degree-based indices
64	ON1V	overall modified Zagreb index of order 1 by valence vertex degrees	Topological indices	Vertex degree-based indices

**Table 4 (continued)**

No.	Name	Description	Block	Sub-Block
65	Qindex	quadratic index	Topological indices	Vertex degree-based indices
66	BBI	Bertz branching index	Topological indices	Vertex degree-based indices
67	DBI	Dragon branching index	Topological indices	Vertex degree-based indices
68	SNar	Narumi simple topological index (log function)	Topological indices	Vertex degree-based indices
69	HNar	Narumi harmonic topological index	Topological indices	Vertex degree-based indices
70	GNar	Narumi geometric topological index	Topological indices	Vertex degree-based indices
71	Xt	total structure connectivity index	Topological indices	Vertex degree-based indices
72	Dz	Pogliani index	Topological indices	Vertex degree-based indices
73	Ram	ramification index	Topological indices	Vertex degree-based indices
74	BLI	Kier benzene-likeness index	Topological indices	Vertex degree-based indices
75	Pol	polarity number	Topological indices	Distance-based indices
76	LPRS	log of product of row sums (PRS)	Topological indices	Distance-based indices
77	MSD	mean square distance index (Balaban)	Topological indices	Distance-based indices
78	SPI	superpendent index	Topological indices	Distance-based indices
79	PJI2	2D Petitjean shape index	Topological indices	Distance-based indices
80	ECC	eccentricity	Topological indices	Distance-based indices
81	AECC	average eccentricity	Topological indices	Distance-based indices
82	DECC	eccentric	Topological indices	Distance-based indices
83	MDDD	mean distance degree deviation	Topological indices	Distance-based indices
84	UNIP	unipolarity	Topological indices	Distance-based indices
85	CENT	centralization	Topological indices	Distance-based indices
86	VAR	variation	Topological indices	Distance-based indices
87	ICR	radial centric information index	Topological indices	Distance-based indices
88	MaxTD	max topological distance	Topological indices	Distance-based indices
89	MeanTD	mean pairwise topological distance	Topological indices	Distance-based indices
90	MaxDD	max detour distance	Topological indices	Distance-based indices
91	MeanDD	mean pairwise detour distance	Topological indices	Distance-based indices
92	SMTI	Schultz Molecular Topological Index (MTI)	Topological indices	MTI indices
93	SMTIV	Schultz Molecular Topological Index by valence vertex degrees	Topological indices	MTI indices
94	GMTI	Gutman Molecular Topological Index	Topological indices	MTI indices
95	GMTIV	Gutman Molecular Topological Index by valence vertex degrees	Topological indices	MTI indices
96	Xu	Xu index	Topological indices	MTI indices
97	CSI	eccentric connectivity index	Topological indices	MTI indices
98	Wap	all-path Wiener index	Topological indices	Path/walk indices
99	S1K	1-path Kier alpha-modified shape index	Topological indices	Path/walk indices
100	S2K	2-path Kier alpha-modified shape index	Topological indices	Path/walk indices
101	S3K	3-path Kier alpha-modified shape index	Topological indices	Path/walk indices
102	PHI	Kier flexibility index	Topological indices	Path/walk indices
103	PW2	path/walk 2 - Randic shape index	Topological indices	Path/walk indices
104	PW3	path/walk 3 - Randic shape index	Topological indices	Path/walk indices
105	PW4	path/walk 4 - Randic shape index	Topological indices	Path/walk indices
106	PW5	path/walk 5 - Randic shape index	Topological indices	Path/walk indices
107	MAXDN	maximal electrotopological negative variation	Topological indices	E-state indices
108	MAXDP	maximal electrotopological positive variation	Topological indices	E-state indices
109	DELS	molecular electrotopological variation	Topological indices	E-state indices
110	TIE	E-state topological parameter	Topological indices	E-state indices
111	Psi_i_s	intrinsic state pseudoconnectivity index - type S	Topological indices	E-state indices
112	Psi_i_A	intrinsic state pseudoconnectivity index - type S average	Topological indices	E-state indices
113	Psi_i_0	intrinsic state pseudoconnectivity index - type 0	Topological indices	E-state indices
114	Psi_i_1	intrinsic state pseudoconnectivity index - type 1	Topological indices	E-state indices
115	Psi_i_t	intrinsic state pseudoconnectivity index - type T	Topological indices	E-state indices
116	Psi_i_0d	intrinsic state pseudoconnectivity index - type 0d	Topological indices	E-state indices
117	Psi_i_1d	intrinsic state pseudoconnectivity index - type 1d	Topological indices	E-state indices
118	Psi_i_1s	intrinsic state pseudoconnectivity index - type 1s	Topological indices	E-state indices
119	Psi_e_A	electrotopological state pseudoconnectivity index - type S average	Topological indices	E-state indices
120	Psi_e_0	electrotopological state pseudoconnectivity index - type 0	Topological indices	E-state indices
121	Psi_e_1	electrotopological state pseudoconnectivity index - type 1	Topological indices	E-state indices
122	Psi_e_t	electrotopological state pseudoconnectivity index - type T	Topological indices	E-state indices
123	Psi_e_0d	electrotopological state pseudoconnectivity index - type 0d	Topological indices	E-state indices
124	Psi_e_1d	electrotopological state pseudoconnectivity index - type 1d	Topological indices	E-state indices
125	Psi_e_1s	electrotopological state pseudoconnectivity index - type 1s	Topological indices	E-state indices
126	BAC	Balaban centric index	Topological indices	Centric indices
127	LOC	lopping centric index	Topological indices	Centric indices

### 3.3 Results

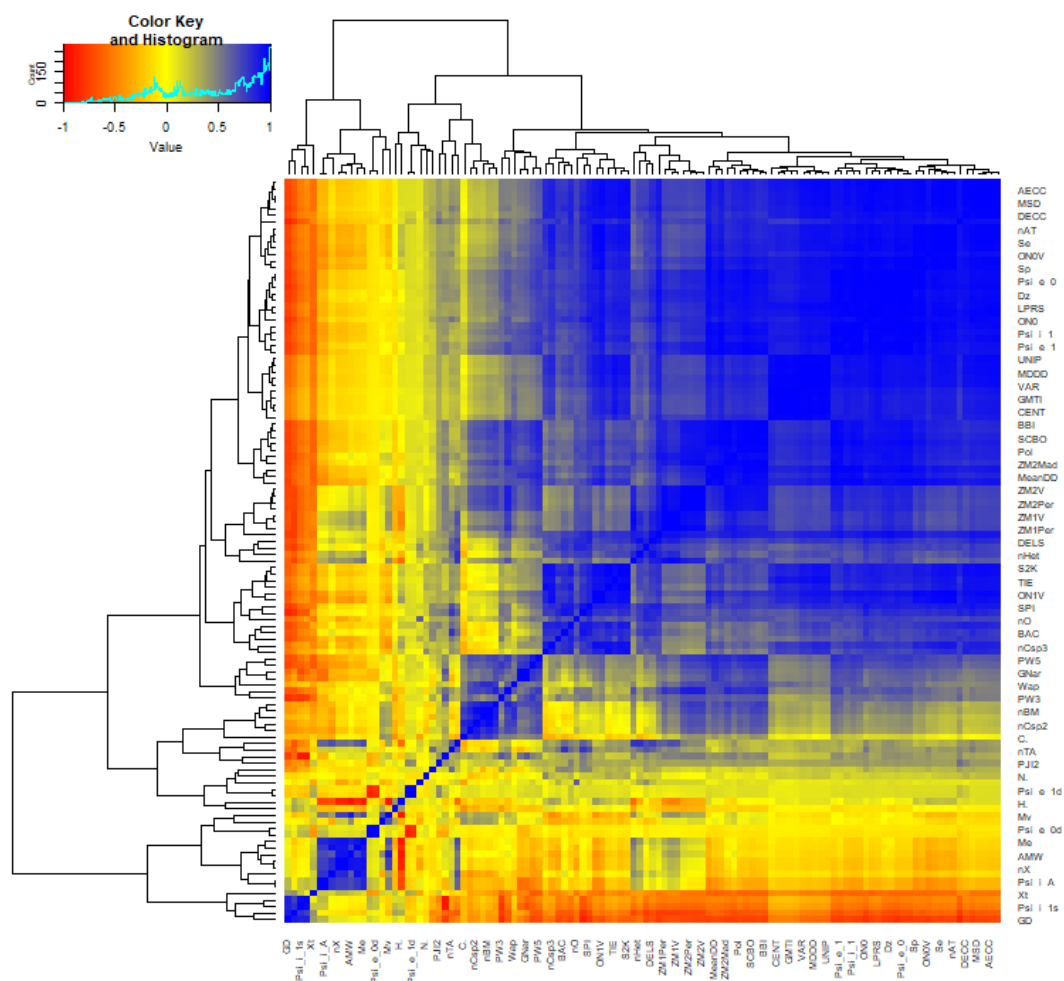
Figure 9 is a heatmap which shows correlation coefficient of the variables. Rows and columns are molecular descriptors. The number of them is 127. Red cell means -1, blue cell means 1, and yellow cell means 0. The descriptors used in this study are based on the 2D structure information. Several descriptors include duplicate information for each other, many of the prepared descriptors contained strong correlations shown in figure 9. In the case of a strong correlation among the explanatory variables called multi-collinearity, ordinary multiple regression analysis cannot be used directly. However, even in this case, Partial Least Squares regressions (PLSR) [85] could be carried out to constructing linear predicting model.

I performed to create mathematical models for relating three polymer properties, i.e., the density ( $\rho$ ), the dissolution parameter ( $\delta$ ), and the glass transition temperature ( $T_g$ ), by 127 descriptors generated from the monomer structure. We tried to predict these three parameters by PLSR. Since the number of data is relatively small (48 data), LOOCV (Leave-One-Out Cross Validation) was applied to the data.

Figures 10 (a-c) show predicted results of PLSR models of each property. Table 4 compares the number of components used in the models, root mean square error (RMSE) and  $R^2$  (determination coefficient) for training data.

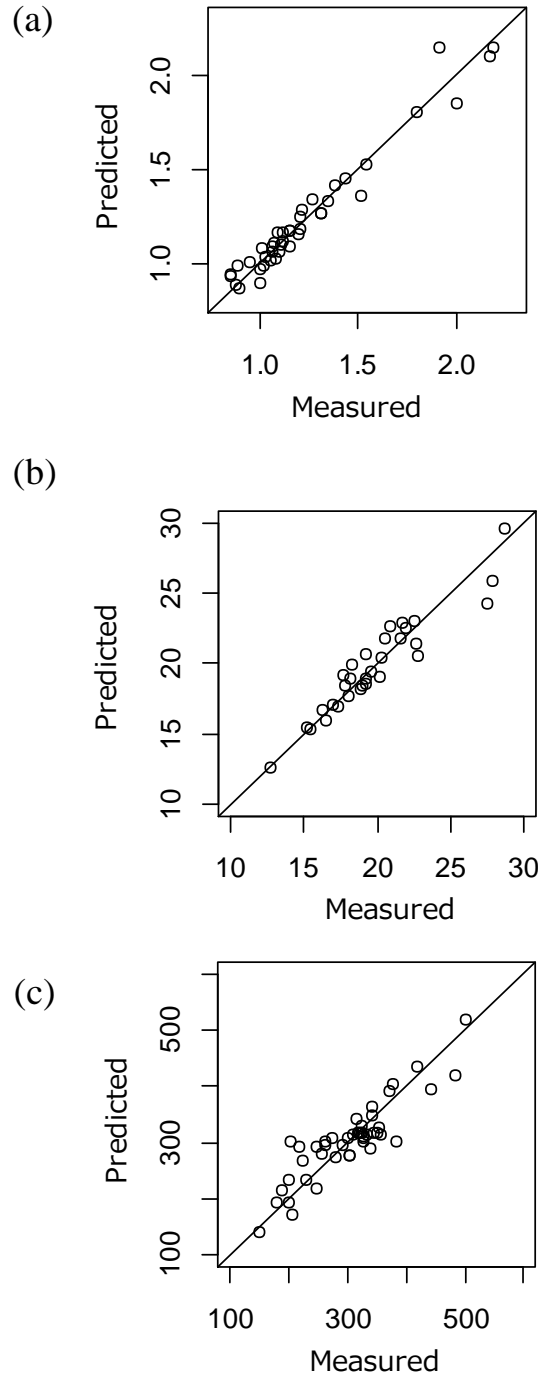
For each model, the value of RMSE went down to a certain point and increased thereafter. In PLSR, it is necessary to determine the number of components to be used in the model. This time, the number of component when contribution rate exceeds 85% for the first time is selected.

$R^2$  for PLSR models were from 0.80 to 0.96. The models could be constructed for these properties and their fitting is good for training dataset, but generalization of prediction model should be evaluated.



**Figure 9: Correlation coefficient of molecular descriptors**

Correlation coefficient of 127 values of molecular descriptor as explanatory variables to predict polymer properties



**Figure 10: The summary of training results of PLSR model ( $\rho$ ,  $\delta$ ,  $T_g$ )**

The result of PLSR model for three properties (a): density, (b): dissolution parameter, (c): glass transition temperature. Predicted values versus measured values were plotted for training data.

**Table 5:** Comparison of PLSR models of three properties

Property	Number of component	RMSE	R <sup>2</sup>
$\rho$	6	0.068	0.96
$\delta$	6	1.1	0.90
$T_g$	5	34	0.80

### 3.4 Discussion

#### Test dataset preparation

For evaluating reliability of models, test dataset which is not concluded in training data is used. Chemical structures of the test data were listed in Table 6. The data were obtained from other source, which is “*Polymer Data Handbook*” edited by James E. Mark [86]. The number of test data is 11. The values of properties (objective variables of regression) were compared to training dataset by scatter plot (Figure 11). The values of descriptors (explanatory variables of regression) were compared to training dataset by hierarchical clustering (Figure 12). These results showed there are no significant outlier data in test dataset, therefore it is appropriate to use this data as evaluating the reliability of obtained models.

#### Considering applicability domain of models

Statistical models have applicability domain (AD). There are some proposal to evaluate AD [88, 89].

In industrial view, it is important to evaluate AD for unknown data. Accurate evaluation of models lead to development of novel materials with predicting their characteristics before synthesize them practically. It is highly contributed to high efficiency performance development. In other words, models which constructed based limited training data should be accurately evaluated for unknown (test) data.

For evaluating AD, T2 statistics and Q statistics which are based on principal component analysis (PCA) can be applied [90]. This method is adequate for dataset with multicollinearity like this case.

$T^2$  statistics are defined by Eq (7).

$$T^2 = \sum_{i=1}^A \left( \frac{t_i}{s_i} \right)^2 \quad (7)$$

Here  $t_i$  means score of  $i$ th principal component (PC),  $s_i$  means standard deviation of  $i$ th PC and  $A$  is the number of PCs need to be considered (for example, it is determined as the component number which firstly gives over 95% contribution ratio).

Q statistics are defined by Eq (8).

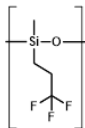

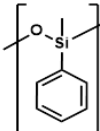
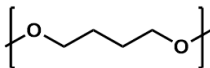
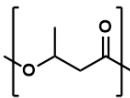


$$Q = \sum_{i=1}^k (y_i - \hat{y}_i)^2 \quad (8)$$

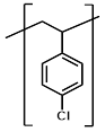
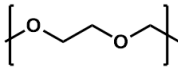
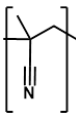
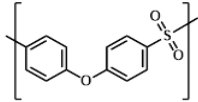
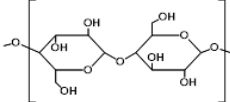
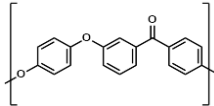
Here  $k$  is the number of variables,  $y_i$  means the value of  $i$ th variable and  $\hat{y}_i$  means estimated value of  $i$ th variable based on PCA (namely inverse mapping of  $i$ th variable using first  $A$ th PCs information).

$T^2$  means the distance of data from the origin based on the information of the first  $A$ th PCs.  $Q$  statistics means the indicator of information which can not be described by the first  $A$ th PCs. When these values get large for test data, the data is outlier from training dataset and it possibly is out of applicability domain. In this study, the datasets (training and test) were analyzed by PCA.  $T^2$  and  $Q$  statistics were calculated which is based on 95% contribution rate. The result are shown in figure 13. Black circles show training data and red show test data. From this plot, four data (TEST 1, 9, 10 and 11) are plotted far from training data and they appear outlier. From this analysis, it is considered that these four data are likely out of AD and prediction results can be worse. As described before, from correlation analysis and hierarchical clustering, no significant difference are appeared about these four data. Using PCA approach you can find abnormal data which cannot detected by simple correlation or clustering analysis. Note that other seven data are considered as applicable data for the prediction model.

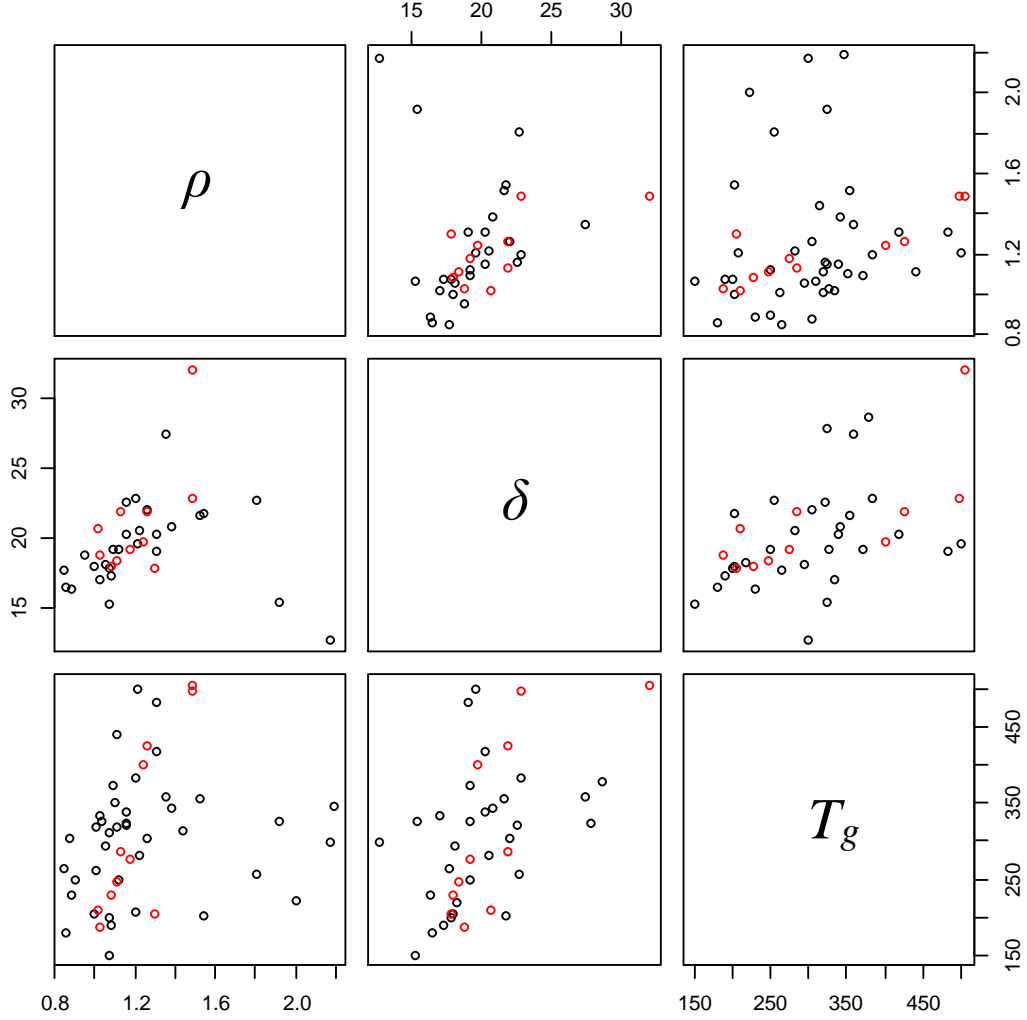
**Table 6:** Chemical structures of test dataset.

No.	Name	Structure
1	Poly(methyltrifluoropropylsiloxane)	
2	Poly(propylene sulfide)	
3	Poly(methylphenylsiloxane)	
4	Poly(1,3-dioxepane)	
5	Poly(hydroxybutyrate)	

---

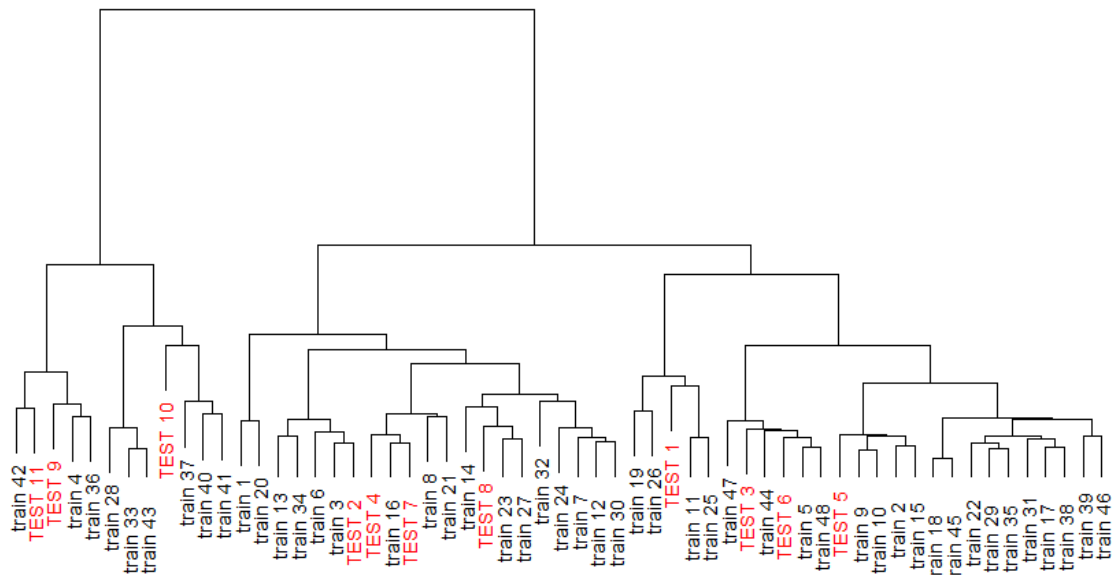
6	Poly(p-chlorostyrene)	
7	Poly(1,3-dioxolane)	
8	Poly(methacrylonitrile)	
9	Poly(ether sulfone)	
10	Cellulose	
11	Poly(ether ether ketone)	

---



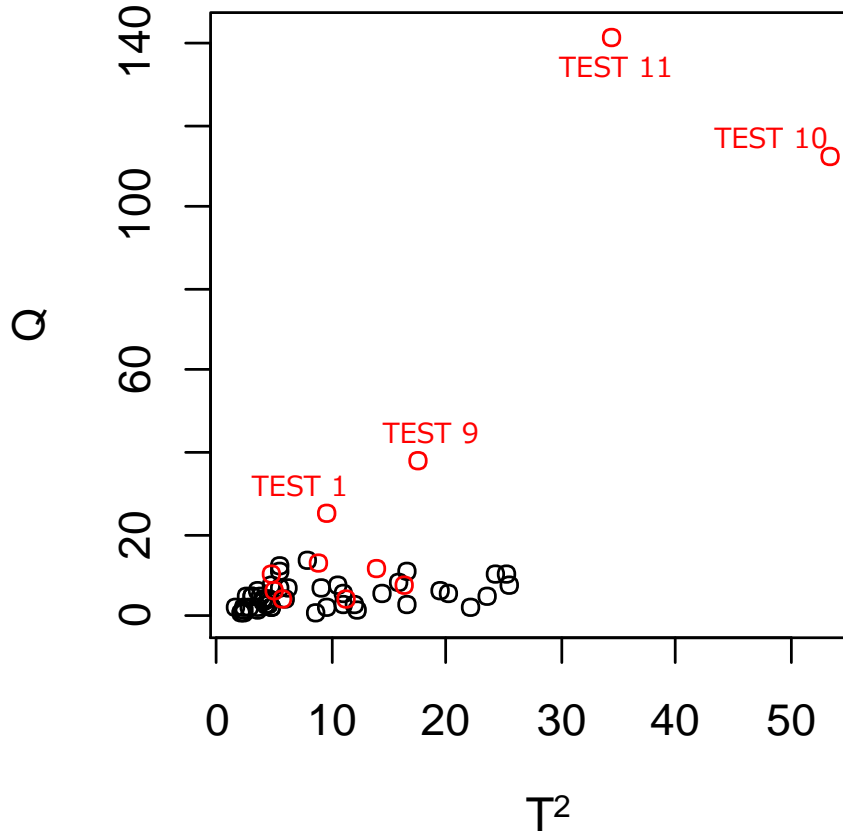
**Figure 11: The scatter plots of training and test dataset ( $\rho$ ,  $\delta$ ,  $T_g$ )**

The scatter plots of training (black) and test (red) dataset. Values of density ( $\rho$ ), solubility parameter ( $\delta$ ) and glass transition temperature ( $T_g$ ) are shown. Range of values of test dataset is similar to training dataset. Correlation coefficients are: 0.10( $\rho$  and  $\delta$ ; 38 data), 0.21( $\rho$  and  $T_g$ ; 50 data), 0.50( $\delta$  and  $T_g$ ; 41 data).



**Figure 12: Hierarchical clustering of training and test dataset ( $\rho$ ,  $\delta$ ,  $T_g$ )**

The result of hierarchical clustering of training (black) and test (red) dataset which is constituted from descriptor values of monomer unit structures. Test data distribute to different clusters of training data. This shows there are no significant outlier data in test dataset.



**Figure 13:  $T^2$  and  $Q$  statistics of training and test dataset based on PCA**

Based on PCA,  $T_2$  and  $Q$  statistics are calculated and plotted. Training dataset (black) and test (red) dataset are shown, and four test data (TEST 1, 9, 10 and 11) appear as outliers.

#### **Results of prediction of test dataset and other consideration**

In Table 7, I compared the root means square errors and coefficient of determinations between only training data and only test data. For  $\rho$ , fitting of test data is good. However, for  $\delta$  and  $T_g$ , the value got much worse than training data only. So models of  $\delta$  and  $T_g$  does not show good accuracy for test data. In other words, these prediction models have poor

reliability for additional data. The plots of prediction result are also shown in figure 14.

In Table 8, more detail result is shown. It summarizes measured values, prediction result and their errors. In figure 15-17, regression results are shown again with labels of the name of test data. In some structures, prediction residuals are likely to large (such as # 1, 2, 3, 4, 8, and 10). The trend can be understood by considering chemical structures of the data. In test dataset, #2: poly-(propylene sulfide) and #9: poly-(ether sulfone) contains S atom. However, in training dataset, there was no structure included S-atom. The difficulty of prediction of #10 (cellulose) came from that training dataset include no glycosidic-bonded structures. #1 (Poly (methyl trifluoro propyl siloxane)) and #3 (Poly (methyl phenyl siloxane)) includes poly-siloxane structure which is included only 1 data in training dataset. Nitrile structure included #8 (Poly-(methacrylonitrile)) were also contained only one data in training dataset.

Generally speaking, it will be concluded that poor reliability of the models comes from the lack of variety of training dataset. However, this situation often happens in real experimental data. Pragmatically, it is helpful to judge the reliability of the model using only already obtained data without any additional test dataset.

In the aspect of using the training dataset effectively, it is helpful that other property data can be used to evaluate reliability of prediction. Based on the training dataset, PLS prediction models were constructed for all properties. The result is summarized in table8. The model accuracies were different depends on property. As I mentioned before, properties can be clustered by hierarchical clustering based on the training dataset. In Table 10, depends on the clusters, prediction results were sorted and average of  $R^2$  for each cluster were calculated.

Figure 18 shows cluster ID and average of  $R^2$  values of each cluster. Please note this clustering came from property information only based on the training dataset not including structural information.

On the other hand, using monomer unit structure information, PLS prediction models have been obtained as shown in chapter1-4. Their fitting for test data was different. For  $\rho$ , fitting of test data was good, however for  $\delta$  and  $T_g$ , fitting of test data were not good. For each cluster, average  $R^2$  value of PLS models for training data are shown in figure17. You can evaluate the reliability of the models from this perspective. For example,  $\rho$  is located in cluster "C". The  $R^2$  value of models of parameters in cluster "C" is high. However, clusters which includes  $\delta$  or  $T_g$  ("F" or "G" each) show low  $R^2$  values. In this way, for each cluster, average  $R^2$  of PLS models for training data relate the reliability of the models.

**Table 7: Comparison of PLSR models of three properties**

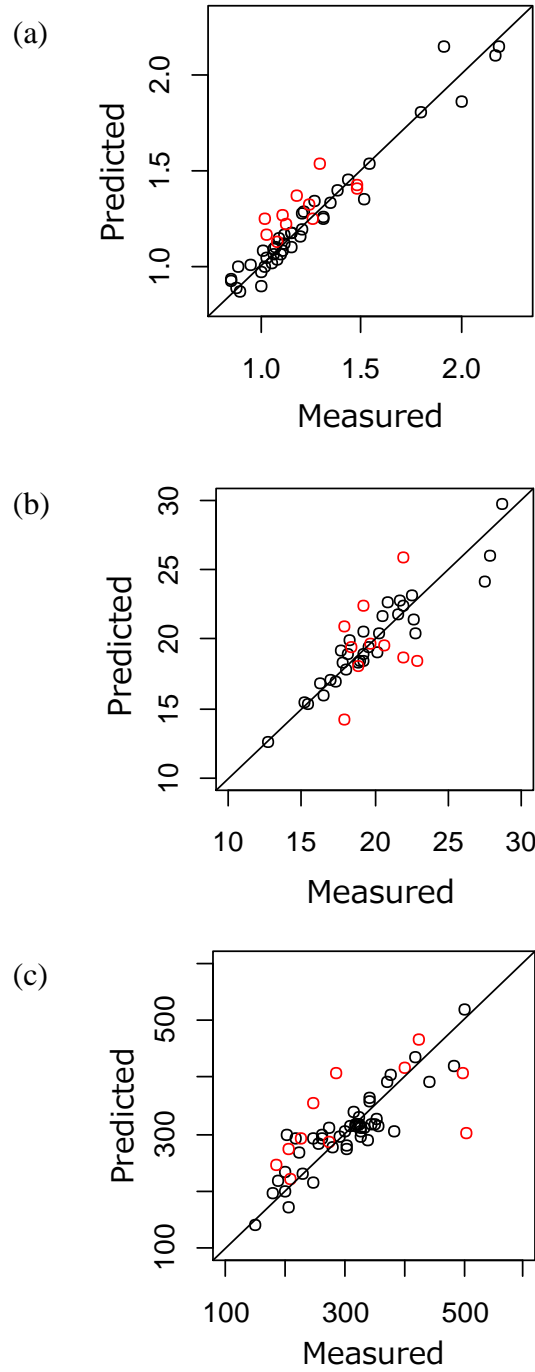
Comparison of root means square errors and coefficient determinations between training and test dataset.

Property	Number of component	RMSE (Training)	R <sup>2</sup> (Training)	RMSE (Test)	R <sup>2</sup> (Test)
$\rho$	6	0.068	0.96	0.14	0.54
$\delta$	6	1.1	0.90	4.2	0.08
$T_g$	5	34	0.80	91	0.39

**Table 8: Prediction result of test dataset**

Error values were calculated as the ratio of prediction residuals per measured value. Gray cells mean absolute of error ratio is large (over 20%).

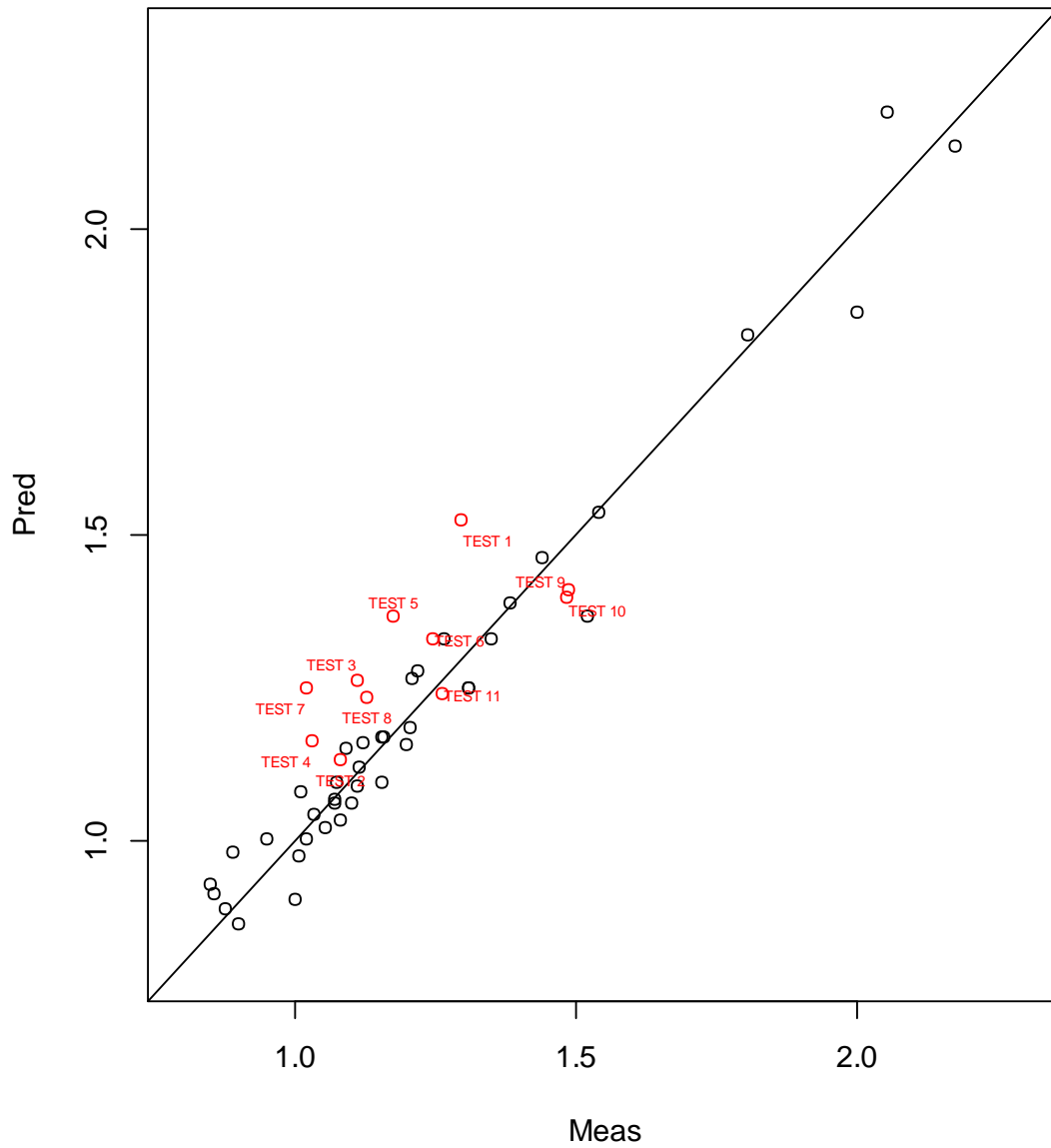
No.	Name	$\rho$			$\delta$			$T_g$		
		Measured	Predicted	Error	Measured	Predicted	Error	Measured	Predicted	Error
1	Poly(methyltrifluoropropylsiloxane)	1.30	1.52	18%	17.9	14.0	-22%	205	266	30%
2	Poly(propylene sulfide)	1.08	1.13	5%	17.9	20.9	17%	228	294	29%
3	Poly(methylphenylsiloxane)	1.11	1.26	14%	18.4	19.4	5%	246	355	44%
4	Poly(1,3-dioxepane)	1.03	1.16	13%	18.8	18.1	-4%	187	247	32%
5	Poly(hydroxybutyrate)	1.18	1.37	16%	19.2	22.4	17%	275	289	5%
6	Poly(p-chlorostyrene)	1.25	1.33	7%	19.7	19.7	0%	400	415	4%
7	Poly(1,3-dioxolane)	1.02	1.25	22%	20.7	19.6	-5%	210	222	5%
8	Poly(methacrylonitrile)	1.13	1.23	9%	21.9	26.0	19%	285	408	43%
9	Poly(ether sulfone)	1.49	1.40	-6%	22.9	18.5	-19%	498	405	-19%
10	Cellulose	1.49	1.41	-5%	32.0	21.6	-33%	505	302	-40%
11	Poly(ether ether ketone)	1.26	1.24	-2%	21.9	18.7	-15%	425	464	9%



**Figure 14: The summary of test results of PLSR model ( $\rho$ ,  $\delta$ ,  $T_g$ )**

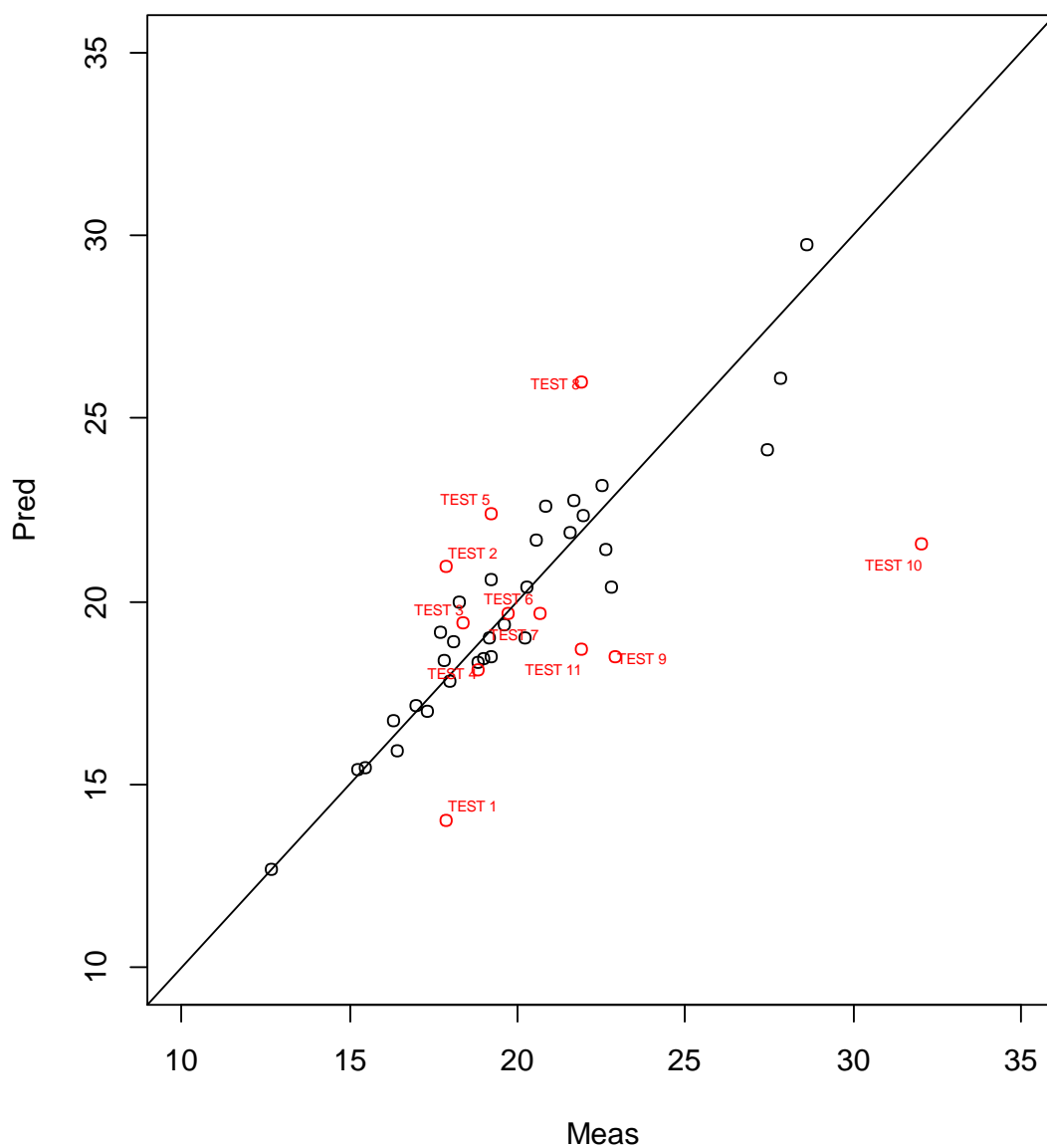
The result of PLSR model for three properties (a): density, (b): dissolution parameter, (c): glass transition temperature. Black points mean training data and red means test data. Predicted values versus measured values were plotted.





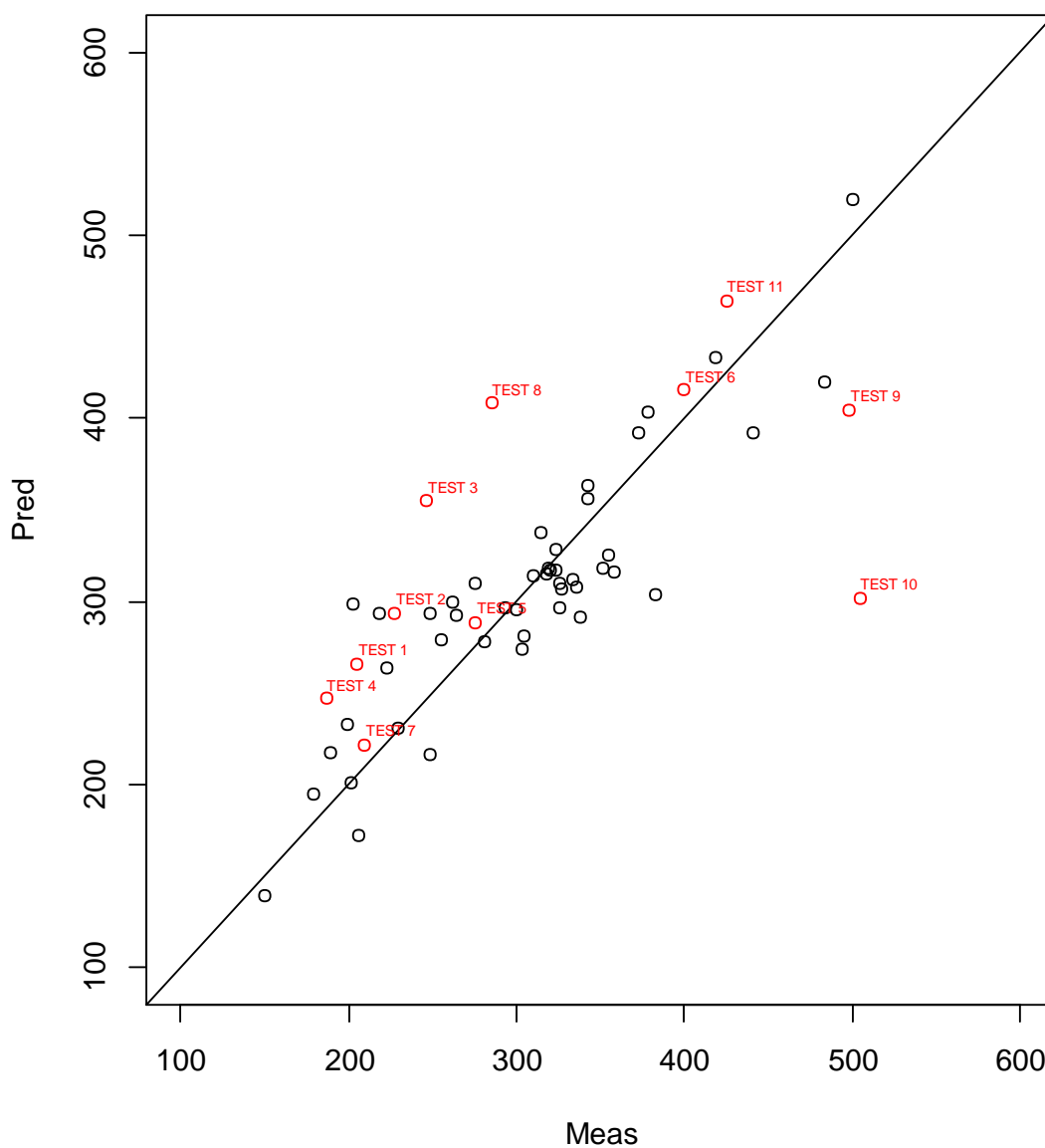
**Figure 15: The test result of PLSR model ( $\rho$ )**

The result of PLSR model for density ( $\rho$ ). Predicted values versus measured values were plotted. Black points mean training data and red means test data. Labels of points indicate the data ID of test data.



**Figure 16: The test result of PLSR model ( $\delta$ )**

The result of PLSR model for solubility parameter ( $\delta$ ). Predicted values versus measured values were plotted. Black points mean training data and red means test data. Labels of points indicate the data ID of test data.



**Figure 17: The test result of PLSR model ( $T_g$ )**

The result of PLSR model for glass transition temperature ( $T_g$ ). Predicted values versus measured values were plotted. Black points mean training data and red means test data. Labels of points indicate the data ID of test data.

**Table 9: Summary of the results of PLS prediction models**

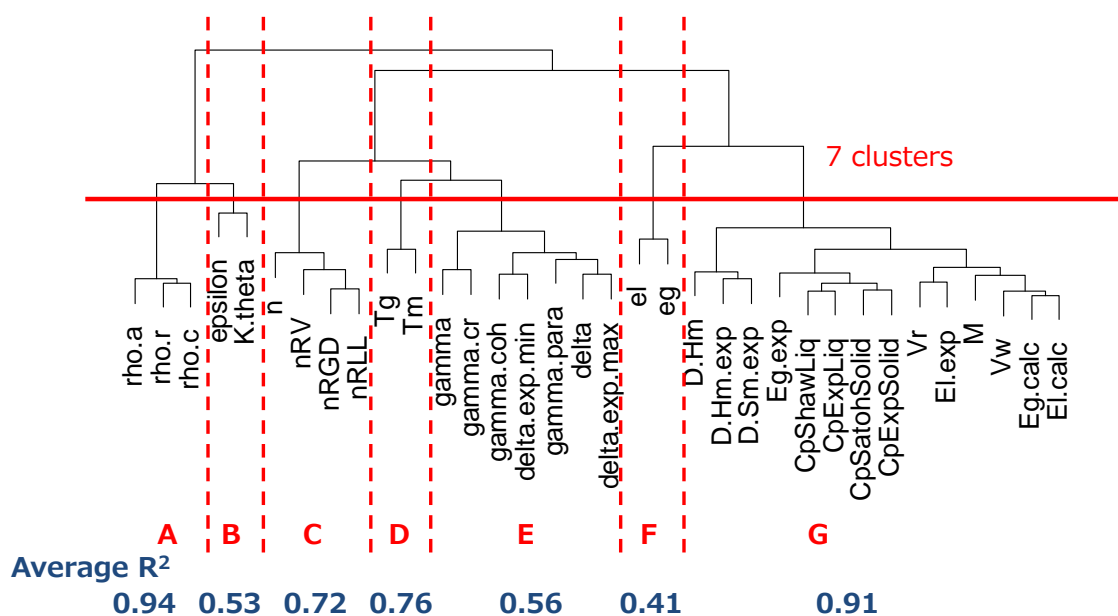
PLS prediction models were constructed for 34 properties, sorted by  $R^2$  of PLS model based on the training dataset.

Property	Number_of_data	PLS_ncomp	PLS_RMSE	PLS_R <sup>2</sup>	Cluster_ID
El.calc	35	8	6.9	1.00	G
Eg.calc	35	7	4.7	1.00	G
CpExpLiq	24	6	4.3	1.00	G
El.exp	31	2	72	0.97	G
rho.a	48	6	0.044	0.97	A
M	48	1	12	0.97	G
Eg.exp	25	3	33	0.97	G
Vw	46	2	8.8	0.96	G
gamma.coh	27	5	2.8	0.92	E
nRLL	24	3	0.019	0.91	C
D.Sm.exp	21	2	8.4	0.91	G
CpExpSolid	30	2	28	0.90	G
CpSatohSolid	32	2	26	0.90	G
rho.c	36	2	0.12	0.89	A
delta.exp.min	26	4	1.2	0.88	E
CpShawLiq	30	2	42	0.88	G
delta	31	5	1.2	0.88	E
rho.r	41	2	0.12	0.87	A
K.theta	24	4	0.027	0.87	B
Vr	41	1	21	0.85	G
D.Hm.exp	21	1	5.9	0.82	G
nRV	24	2	0.021	0.81	C
nRGD	24	2	0.028	0.81	C
Tg	47	5	34	0.80	D
Tm	39	3	60	0.67	D
D.Hm	30	1	9.4	0.64	G
delta.exp.max	23	1	2.8	0.46	E
gamma.para	30	2	6.9	0.41	E
el	33	1	1.1	0.40	F
eg	31	1	0.59	0.40	F
n	36	1	0.054	0.35	C
gamma.cr	29	1	6.2	0.18	E
gamma	32	1	5.8	0.13	E
epsilon	31	1	1.8	0.07	B

**Table 10: Summary of the results of PLS prediction models (sorted)**

Summary of the results of PLS prediction models were constructed for 34 properties, sorted by the clustering result on the training dataset.

Property	Number_of_data	PLS_ncomp	PLS_RMSE	PLS_R <sup>2</sup>	Cluster_ID	Average of R <sup>2</sup> for the cluster
rho.a	48	6	0.044	0.97	A	0.91
rho.c	36	2	0.12	0.89		
rho.r	41	2	0.12	0.87		
K.theta	24	4	0.027	0.87	B	0.47
epsilon	31	1	1.8	0.07		
nRLL	24	3	0.019	0.91	C	0.72
nRV	24	2	0.021	0.81		
nRGD	24	2	0.028	0.81		
n	36	1	0.054	0.35	D	0.73
Tg	47	5	34	0.80		
Tm	39	3	60	0.67		
gamma.coh	27	5	2.8	0.92	E	0.55
delta.exp.min	26	4	1.2	0.88		
delta	31	5	1.2	0.88		
delta.exp.max	23	1	2.8	0.46	F	0.40
gamma.para	30	2	6.9	0.41		
gamma.cr	29	1	6.2	0.18		
gamma	32	1	5.8	0.13	F	0.40
el	33	1	1.1	0.40		
eg	31	1	0.59	0.40		
El.calc	35	8	6.9	1.00	G	0.91
Eg.calc	35	7	4.7	1.00		
CpExpLiq	24	6	4.3	1.00		
El.exp	31	2	72	0.97		
M	48	1	12	0.97		
Eg.exp	25	3	33	0.97		
Vw	46	2	8.8	0.96		
D.Sm.exp	21	2	8.4	0.91		
CpExpSolid	30	2	28	0.90		
CpSatohSolid	32	2	26	0.90		
CpShawLiq	30	2	42	0.88		
Vr	41	1	21	0.85		
D.Hm.exp	21	1	5.9	0.82		
D.Hm	30	1	9.4	0.64		



**Figure 18: Comparison of clustering result and PLSR prediction models  $R^2$**

Relationships of hierarchical clustering and generalization performance of PLSR prediction models. Result of hierarchical clustering of properties. 7 clusters are obtained and they are named as A-G. PLSR models were constructed to all properties and their average  $R^2$  values of regression result were shown to each clusters.

## **4. Outlook of materials informatics in the future**

### **4.1 Trend of big database in chemical field**

In recent years, efforts have been made in the chemical industry to collect and provide large chemical databases. There are many available databases [92-97]. This situation will encourage the use of data. It could widen the gap between those who have data and those who don't. However, even if you have a lot of data, the quality and usage of the data is still important. If you can't use data correctly, the results from it will be incorrect. The characteristics and limitations of the data must be properly understood.

### **4.2 The limitation of data usage these days and importance of methodology to interpret data**

In the field of polymer material development, databases are inevitably incomplete. For example, values of the glass transition temperature and viscosity of polymers vary depending on the measurement conditions (temperature and measurement accuracy). It is impossible to apply totally the same experimental conditions to all polymers. Practically, when it comes to develop polymer materials, performance values based on specific evaluation conditions are always required, so training data always remain small. Therefore, as discussed in this paper, the approach to gain helpful knowledge by dealing with the data at hand will become even more important in the future.

There is big trend in materials informatics that Bayesian optimization is powerful tool to discover useful material structure or best formulation of the product [98-101]. It is an exploratory approach, which is effective method to discover optimized chemical compositions when field to explore were set clearly. However, another approach is needed for interpret obtained data already. Pragmatically obtained dataset is always a small part of whole chemical field to be explored. In that situation, the methodology I proposed in this work is able to reveal the concept how to use meaningfully small dataset already obtained. Namely I showed how unsupervised manner apply to classify or understand polymer structures based on small and imperfect dataset. When prediction model is constructed for one target property, other property dataset will be effectively used as evaluation of the generalization applicability of the prediction model. This methodology enables to engineers who work develop novel material product using data-science approach because they have evidence for the performance of data-driven prediction. The engineers will be able to rely on the models so that this leads to increase the success rate

of materials-informatics project. Generally speaking, when introducing a new method, it is critical factor to the success how much you trust in the method.



## 5. Conclusive remarks

Data of 48 types of polymers consisting of 34 properties were prepared based on the literature [34]. The relationship between polymer type and properties was analyzed by hierarchical clustering. Several properties are reflected by the structure of the monomer unit. For these polymers, descriptors were calculated from the 2D SMILES information of the monomer units and the physical properties were estimated. The prediction model by PLSR was obtained for three different property values. Reliability of obtained models could be evaluated using clustering result based on training dataset.

Data-driven approaches, in other words, accumulation of molecular data concerning polymers and selection of optimal models for predicting individual polymer properties are needed strongly in development novel polymer materials fundamentally. Furthermore, it will become important approach for industry to use incomplete dataset to make helpful prediction model and to evaluate it in the future. These approach clear practical problems and realize applying data-science techniques to industrial chemical data.

## Research achievement

- The contents of chapter 2-1, 2-2 and 3-1 were published as: Relationships of Polymer Properties and Predicting Polymer Properties Based on Monomer Structure Information, Hitoshi Yamano, Tomoyuki Miyao, Naoaki Ono, Aki Morita, Shigehiko Kanaya, *Journal of Computer Aided Chemistry*, Vol.20, 84-91 (2019).
- The contents of chapter 2-1, 2-2 and 3-1 were presented as: Considering and predicting properties of general polymers based on their monomer unit structure, Hitoshi Yamano, Tomoyuki Miyao, Naoaki Ono, Aki Morita, Shigehiko Kanaya, 6th Autumn School of Chemoinformatics (poster session) in Nara, Japan (2019).
- The contents of chapter 3-1, 3-2 and 3-3 were presented as: Clustering and predicting properties of general polymers based on their monomer unit structure, Hitoshi Yamano, Hiroaki Shimizu, Shigehiko Kanaya, Tomoyuki Miyao, Aki Morita, Naoaki Ono, American Chemical Society Fall 2020 Virtual Meeting & Expo, CINF-113.
- The contents of chapter 3-1, 3-2 and 3-3 were presented as: Predicting and considering properties of general polymers using incomplete dataset, Hitoshi Yamano, Hiroaki Shimizu, Shigehiko Kanaya, Tomoyuki Miyao, Aki Morita, Naoaki Ono, International Symposium on Semiconductor Manufacturing Virtual Symposium 2020, MI-036.

## Acknowledgement

I would like to express my deepest gratitude to Professor Kanaya and all the members of Kanaya Laboratory in Nara Institute of Science and Technology. I deeply thank to my supervisors and co-supervisors for their valuable lessons. In particular, I would like to thank Prof. Kanaya, Assoc. Prof. Ono, and Assoc. Prof. Miyao for their kind and detailed explanations, which were very helpful for me. I would like to thank Ms. Morita and students in Kanaya Lab. for their great cooperation in the preparation of the data.

I would like to express my deep gratitude to the people of Tokyo Ohka Kogyo Co., Ltd. for giving me the opportunity to do this research. I would like to thank Mr. Harutoshi Sato, Mr. Katsumi Ohmori, and Mr. Hiroataka Yamamoto for allowing one employee to do this research. I deeply thank to Mr. Hiroaki Shimizu, Dr. Ryohei Eguchi, and the other members of the AI Promotion Section, who arranged their work for me regarding the time spent on this research and gave their extensive research advice.

I would like to thank Prof. Masakazu Iwamoto and Dr. Takashi Deguchi who are supervisors during my graduate school days for their teachings.

Finally, I thank God, my parents for raising me, my daughters for supporting me, and my wife.

## References

1. T. Hey, S. Tansley, and K. Tolle (Eds.), *The Fourth Paradigm -Data-Intensive Scientific Discovery*, Microsoft Research (2009)
2. A. Smith, M. Gerstein, D. J. Weitzner, T. B.-Lee, W. Hall, J. Hendler, N. Shadbolt and D. J. Weitzner, *Science*, **314**, 1682 (2006)
3. J. Dean, arXiv:1911.05289 (2019)
4. R. Inayama and K. Kadowaki, *AIJ J. Technol. Des.* **25**, 863-867 (2019)
5. A. Mirhoseini, A. Goldie, M. Yazgan, J. Jiang, E. Songhori, S. Wang, Y.-J. Lee, E. Johnson, O. Pathak, S. Bae, A. Nazi, J. Pak, A. Tong, K. Srinivasa, W. Hang, E. Tuncer, A. Babu, Q. V. Le, J. Laudon, R. Ho, R. Carpenter and J. Dean, arXiv:2004.10746 (2020)
6. H. Nishiuchi, M. Taguri and Y. Ishikawa, *PLoS ONE*, **11** (7): e0158328
7. Y. Iwasawa, I. E. Yairi and Y. Matsuo, *Transactions of the Japanese Society for Artificial Intelligence*, **32**, 1-12 (2017)
8. T. Okazaki, K. Okusa and K. Yoshida, *2018 International Symposium on Semiconductor Manufacturing (ISSM)*, Tokyo, Japan, 2018, 1-3, doi: 10.1109/ISSM.2018.8651135.
9. S. R. Kalidindi and M. D. Graef, *Annual Review of Materials Research*, **45**, 171-193 (2015)
10. S. Kalinin, B. Sumpter and R. Archibald, *Nature Mater.*, **14**, 973–980 (2015)
11. Y. Iwasaki, M. Ishida and M. Shirane, *Sci. Technol. Adv. Mater.*, **21**, 25–28 (2020)
12. C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chemistry of Materials*, **31**, 3564-3572 (2019)
13. J. P. Reid and M.S. Sigmon, *Nature*, **571**, 343 (2019)
14. S. Wu, G. Lambard, C. Liu, H. Yamada and R. Yoshida, *Molecular Informatics*, **39**, 1900107 (2020)
15. M. Kano and Y. Nakagawa, *Comput. Chem. Eng.*, **32**, 12-24 (2008)
16. P. Kadlec, B. Gabrys and S. Strandt, *Comput. Chem. Eng.*, **33**, 795-814 (2009)
17. H. Kaneko and K. Funatsu, *Chemometrics and Intelligent Laboratory Systems*, **153**, 75-81 (2016)
18. H. Kaneko and K. Funatsu, *Ind. Eng. Chem. Res.*, **54**, 12630–12638 (2015)
19. S. Takeda, H. Kaneko, and K. Funatsu, *Journal of Chemical Information and Modeling*, **56**, 1885-1893 (2016)
20. R. Eguchi, N. Ono, A. H. Morita, T. Katsuragi, S. Nakamura, M. Huang, Md

- Altaf-Ul-Amin, S. Kanaya, *BMC Bioinformatics*, **20**, 380 (2019)
21. L. Ward and C. Wolverton, *Current Opinion in Solid State and Materials Science*, **21**, 167-176 (2017)
  22. J. Behler, *J. Chem. Phys.* **145**, 170901 (2016)
  23. A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, P. G. Debenedetti, *Chemical Physics Letters*, **509**, 1–11 (2011)
  24. A. W. Long and A. L. Ferguson, *J. Phys. Chem. B*, **118**, 4228–4244 (2014)
  25. P. Hajji, L. David, J. F. Gerard, J. P. Pascault and G. Vigier, *Journal of Polymer Science Part B Polymer Physics*, **37**, 3172 (1999)
  26. P. M. Hergenrother, *High Performance Polymers*, **15**, 3 (2003)
  27. R. Pó and N. Cardi, *Progress in Polymer Science*, **21**, 47-88 (1996)
  28. U. Ali, K. Juhanni, B. A. Karim and N. A. Buang, *Journal, Polymer Reviews*, **55**, 678 (2015)
  29. K. Satoh, *The Material Constants Estimation Method*, Maruzen Publishing (1977)
  30. A. Fredenslund, R. L. Jones and J. M. Prausnitz, *AIChE Journal*, **21**, 1086-1099 (1975)
  31. T. Yamamoto and H. Furukawa, *Kobunshi Ronbunshu* (in Japanese), **52**, 187-193 (1995)
  32. T. Suzuki and M. Ishida, *Journal of the Fuel Society of Japan* **61**, 383-389, (1982)
  33. S. Ando, *Kogaku* (in Japanese), **44**, 298 (2015)
  34. D.W. V. Krevelen, *Properties of Polymers. 4th Edition*, Elsevier (2009)
  35. J. Bicerano, *Prediction of Polymer Properties, 3rd edition*, CRC Press (2002)
  36. K. Nagasaka, *Kobunshi* (in Japanese), **40**, 744-747 (1991)
  37. T. M. Madkour and A.M. Barakat, *Comput. and Theoretical Polymer Science*, **7**, 35-46 (1997)
  38. Y. Y. Gotlib, N. K. Balabaev, A. A. Darinski, and I. M. Neelov, *Macromolecules*, **13**, 602 (1980)
  39. M. Bishop, M. H. Kalos, and H. L. Frisch, *J. Chem. Phys.*, **70**, 1299 (1979)
  40. G. S. Grest and K. Kremer, *Physical Review A*, **33**, 3628 (1986)
  41. Y. Natsume, T. Minakata and T. Aoyagi, *Organic Electronics*, **10**, 107-114 (2009)
  42. T. Aoyagi, *Journal of the Society of Rheology, Japan*, **37**, 75-79 (2009)
  43. K. Tagashira, K. Z. Takahashi and T. Aoyagi, *Materials*, **11**, 83 (2018)
  44. H. Miyamoto, M. Umemura, T. Aoyagi, C. Yamane, K. Ueda and K. Takahashi,

- Carbohydrate Research*, **344**, 1085-1094 (2019)
45. R. Garci'a-Domenech and J. V. de Julia'n-Ortiz, *J. Phys. Chem. B*, **106**, 1501-1507 (2002)
  46. A. R. Katritzky, P. Rachwal, K. W. Law, M. Karelson, and V. S. Lobanov, *J. Chem. Inf. Comput. Sci.*, **36**, 879-884 (1996)
  47. C. Duce, A. Micheli, A. Starita, M. R. Tine' and R. Solaro, *Macromol. Rapid Commun.* **27**, 711-715 (2006)
  48. A. Liu, X. Wang, L. Wang, H. Wang and H. Wang, *European Polymer Journal*, **43**, 989-995 (2007)
  49. B. E. Mattioni and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, **42**, 232-240 (2002)
  50. D. Ajloo, A. Sharifian and H. Behniafar, *Bull. Korean Chem. Soc.*, **29**, 2009-2016 (2008)
  51. A. J. Holder, L. Ye, J. D. Eick and C. C. Chappelow, *QSAR Comb. Sci.*, **25**, 905-911 (2006)
  52. A. Afantitis, G. Melagraki, H. Sarimveis, P. A. Koutentis, J. Markopoulos and O. Igglessi-Markopoulou, *Polymer*, **47**, 3240-3248 (2006)
  53. C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, R. Ramprasad, *J. Phys. Chem. C*, **122**, 17575-17585 (2018)
  54. T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, R. Ramprasad, *Sci. Data*, **3** 160012 (2016)
  55. Mannodi-Kanakkithodi, A. Chandrasekaran, C. Kim, T. D. Huan, G. Pilania, V. Botu, R. Ramprasad, *Materials Today* **21**, 785-796 (2018)
  56. Mannodi-Kanakkithodi, G. M. Treich, T. D. Huan, R. Ma, M. Tefferi, Y. Cao, G. A. Sotzing and R. Ramprasad, *Adv. Mater.* **28**, 6277 (2016)
  57. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman and R. Ramprasad, *Sci. Rep.* **6**, 20952 (2016)
  58. J. P. Lightstone, L. Chen, C. Kim, R. Batra and R. Ramprasad, *J. Appl. Phys.*, **127**, 215105 (2020)
  59. Jha, A. Chandrasekaran, C. Kim and R. Ramprasad, *Modelling Simul. Mater. Sci. Eng.* **27**, 024002 (2019)
  60. A. Chandrasekaran, C. Kim, S. Venkatram and R. Ramprasad, *Macromolecules*, **53**, 4764 (2020)
  61. S. Venkatram, C. Kim, A. Chandrasekaran and R. Ramprasad, *J. Chem. Inf. Model.*, **59**, 4188-4194 (2019)
  62. T. D. Huan, A. Mannodi-Kanakkithodi and R. Ramprasad, *Phys. Rev. B*, **92**, 014106 (2015)

63. H. D. Tran, C. Kim, L. Chen, A. Chandrasekaran, R. Batra, S. Venkatram, D. Kamal, J. P. Lightstone, R. Gurnani, P. Shetty, M. Ramprasad, J. Laws, M. Shelton and R. Ramprasad, *Journal of Applied Physics*, **128**, 171104 (2020)
64. K. Hatakeyama-Sato, T. Tezuka, Y. Nishikitani, H. Nishide, and K. Oyaizu, *Chem. Lett.*, **48**, 130–132 (2019)
65. H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa and R. Yoshida, *ACS Cent. Sci.*, **5**, 1717–1730 (2019)
66. National Institute for Materials Science (NIMS) Press release, 2019.06.26 (<https://www.nims.go.jp/eng/news/press/2019/06/201906260.html>)
67. S. Wu, Y. Kondo, M. 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*, **5**, 66 (2019)
68. J. J. de Pablo, B. Jones, C. L. Kovacs, V. Ozolins and A. P. Ramirez, *Current Opinion in Solid State and Materials Science*, **18**, 99-117 (2014)
69. J. J. de Pablo, N. E. Jackson, M. A. Webb, Long-Qing Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. DeLongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton and Ji-Cheng Zhao, *Comput Mater.*, **5**, 41 (2019)
70. S. Yagyu, M. Yoshitake and T. Chikyo, *Vacuum and Surface Science*, **61**, 196-199 (2018) (in Japanese)
71. H. Koinuma and I. Takeuchi, *Nature Materials*, **3**, 429 - 438 (2004)
72. M. L. Green, I. Takeuchi, and J. R. Hattrick-Simpers, *J. Appl. Phys.*, **113**, 231101 (2013)
73. T. Nagata, Y. Suzuki, Y. Yamashita, A. Ogura and T. Chikyow, *Jpn. J. Appl. Phys.* **57**, 04FJ04 (2018)
74. T. Miyao, M. Arakawa and K. Funatsu, *Mol. Inf.*, **29**, 111-125 (2010)
75. R. Yoshida, NIMS MI2I final report (in Japanese) (<https://www.nims.go.jp/MI2I/event/d53p8f000000a9j0-att/d53p8f000000d3sz.pdf>) (2020)
76. S. Ju, R. Yoshida, C. Liu, K. Hongo, T. Tadano and J. Shiomi, *ChemRxiv Preprint*. <https://doi.org/10.26434/chemrxiv.9850301.v1> (2019)
77. H. Numazawa, Y. Igarashi, K. Sato, H. Imai, Y. Oaki, *Adv. Theory Simul.*, **2**, 1900130 (2019)
78. Stephen Wu, Yukiko Kondo, Masa-aki Kakimoto, Bin Yang, Hironao Yamada, Isao Kuwajima, Guillaume Lambard, Kenta Hongo, Yibin Xu, Junichiro Shiomi, Christoph Schick, Junko Morikawa and Ryo Yoshida, *npj Comput Mater*. **5**, 66

- (2019)
79. J. G. Hautier, S. P. Ong and K. Persson, *Journal of Materials Research*, **31**, 977-994 (2016)
  80. M. Kobayashi, *Materials Stage* (in Japanese), **20**, 42-48 (2020)
  81. S. Mukaida, *Materials Stage* (in Japanese), **20**, 49-54 (2020)
  82. J. H. Ward Jr., *J. Am. Stat. Assoc.*, **58** (1963)
  83. D. Weininger, *J. Chem. Inf. Comput. Sci.*, **28**, 31-36 (1988)
  84. Alvascience, alvaDesc (software for molecular descriptors calculation) version 1.0.16, 2019, <https://www.alvascience.com>
  85. S. Wold, M. Sjostrom and L. Eriksson, *Chemom. Intell. Lab. Syst.*, **58**, 109–130 (2001)
  86. James E. Mark (Ed.), *Polymer Data Handbook*, Oxford University Press (1999)
  87. S. Kim, *Encyclopedia of Bioinformatics and Computational Biology*, **2**, 628-639 (2019)
  88. H. Kaneko, M. Arakawa, K. Funatsu, *AIChE J.*, **57**, 1506 (2011)
  89. Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J, Mekenyan O, *J. Chem. Inf. Model.*, **45**, 839-849 (2005)
  90. T. Kourti, J. F. MacGregor, *Chemometrics and Intelligent Laboratory Systems*, **28**, 3-21 (1995)
  91. M. Nakata, T. Shimazaki, *Journal of Chemical Information and Modeling*, **57**, 1300-1308 (2017)
  92. NIMS press release “NIMS Inorganic Material Database “AtomWork-Adv Made Available to the Public (Fee-Based Service)” (2018)  
(<https://www.nims.go.jp/eng/news/press/2018/05/201805150.html>)
  93. “MI2I know-how report” (in Japanese) (2019)  
(<https://www.jst.go.jp/ihub/files/seika-nims.pdf>)
  94. The PLUMED consortium, *Nature Methods*, **16**, 667–673 (2019)
  95. Balazs Pejo, arXiv:2007.06236 [cs.LG] (2020)
  96. B. K. Wheatle, E. F. Fuentes, N. A. Lynd, and V. Ganesan, *Macromolecules*, **53**, 9449–9459 (2020)
  97. K. Nakano, Y. Noda, N. Tanibata, H. Takeda, M. Nakayama, R. Kobayashi and I. Takeuchi, *APL Materials* **8**, 041112 (2020)
  98. T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi and K. Tsuda, *Materials Discovery*, **4**, 18-21 (2016)
  99. T. Minami and Y. Okuno, *MRS Advances*, **3**, 2975-2980 (2018)



- 100.J. S. Tokarski, A.J. Hopfinger, J. D. Hobbs, D. M. Ford, Jean-Loup M. Faulon, *Computational and Theoretical Polymer Science*, **7**, 199-214 (1997)
- 101.Tu Le, V. Chandana Epa, Frank R. Burden, and David A. Winkler, *Chemical Reviews*, **112**, 2889-2919 (2012)