Taylor & Francis
Taylor & Francis Group

Science and Technology of Advanced Materials: Methods

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

NIMS polymer database PoLyInfo (I): an overarching view of half a million data points

Masashi Ishii, Takuro Ito, Hiroko Sado & Isao Kuwajima

To cite this article: Masashi Ishii, Takuro Ito, Hiroko Sado & Isao Kuwajima (2024) NIMS polymer database PoLyInfo (I): an overarching view of half a million data points, Science and Technology of Advanced Materials: Methods, 4:1, 2354649, DOI: 10.1080/27660400.2024.2354649

To link to this article: https://doi.org/10.1080/27660400.2024.2354649

Published online: 10 Jul 2024.

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2024, VOL. 4, NO. 1, 2354649
https://doi.org/10.1080/27660400.2024.2354649

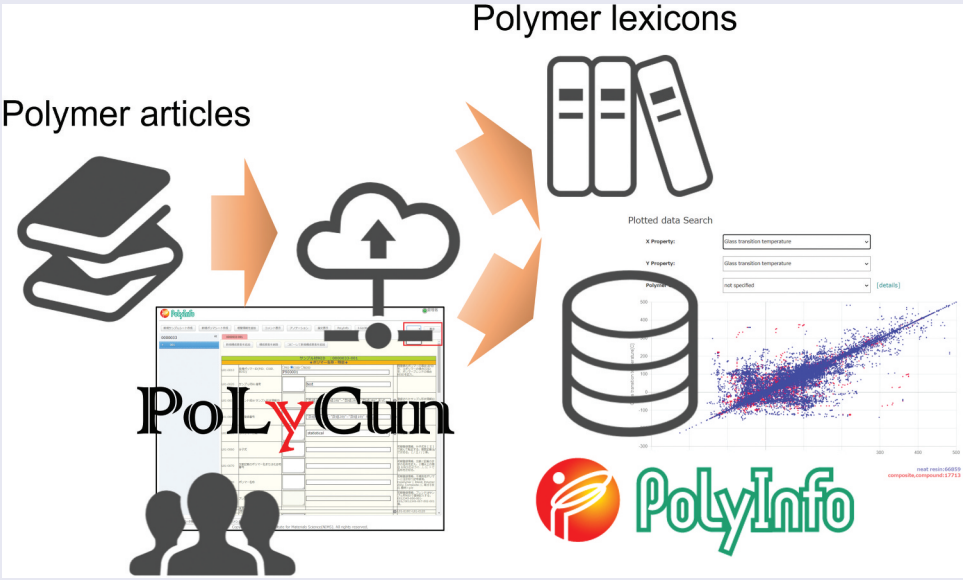
NIMS polymer database PoLyInfo (I): an overarching view of half a million data points

Masashi Ishii[a], Takuro Ito[b], Hiroko Sado[b] and Isao Kuwajima[b]

[a]Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba, Japan; [b]Research Network and Facility Services Division, National Institute for Materials Science (NIMS), Tsukuba, Japan

ABSTRACT

This paper reviews PoLyInfo, which is a polymer database of National Institute for Materials Science of Japan, containing more than half a million data points, from the perspective of a data editor. In particular, we describe how human-readable information provided by a graphical user interface (GUI) is curated, compiled, and displayed on the GUI, and how it can be used for data retrieval. We also describe the data curation policy of PoLyInfo, which can be observed through search functions and data tables, and show the unique taxonomy of various polymers. The status of polymers and their potential and limitations, as seen through a large dataset, are discussed. With this information, we introduce the capabilities of PoLyInfo and provide guidelines for its use as a general database for polymer chemistry, as well as notes on newer data applications such as machine learning.



IMPACT STATEMENT

The NIMS polymer database PoLyInfo created over half a million data and a precise data points architecture through over 20 years of continuous manual data extraction and polymer structure lexicography.

ARTICLE HISTORY
Received 26 March 2024
Revised 30 April 2024
Accepted 6 May 2024

KEYWORDS
Polymer; database; PoLyInfo; human-readable; data curation; lexicon; structure determination; data search; graphical user interface

1. Introduction

PoLyInfo [1], which is a polymer database managed by National Institute for Materials Science (NIMS), contains more than half a million polymer data points manually collected over several decades. Originally expected to be used as a reference for polymer chemistry, this database is now being employed in a variety of unprecedented applications, including as training data for machine learning and material exploration in unexplored areas, as data-driven science is becoming increasingly popular [2,3]. Because of the large amount of accumulated data and the difficulty in verifying them all, if users do not understand the data policy and structure of PoLyInfo, they may chemically misuse the data, affect the quality of secondary data, or waste time in material development by building and using inappropriate prediction models. To appropriately use PoLyInfo and preserve it as an

CONTACT Masashi Ishii ISHII.Masashi@nims.go.jp Center for Basic Research on Materials, National Institute for Materials Science (NIMS), 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan

asset for future generations, we need an overview of the data that can help determine what is good and what is lacking in PoLyInfo, and what must be done to continue to maintain and utilize polymer data. This has been the motivation for writing a series of papers on PoLyInfo, one of which is presented herein. For ease of understanding, the first paper (hereinafter referred to as PoLyInfo (I)) discusses the following topics and provides the core knowledge necessary for understanding the second paper, PoLyInfo (II):

1. Introduction
2. Outline of PoLyInfo
3. Sample ID system
4. Search functions in graphical user interface
5. Database contents
    - 5.1. Master information
    - 5.2. Material information
    - 5.3. Fabrication information
    - 5.4. Formation information
    - 5.5. Properties information
6. Challenges and prospects
7. Conclusion

PoLyInfo (I) provides a human-readable view of the data, whereas PoLyInfo (II) provides a machine-readable view of polymeric concepts and data structures in an ontological form. The two papers largely cover the same topics and aim to treat human- and machine-readable polymer knowledge equally.

Some of the major organic chemistry databases mainly containing small molecular data are SciFinder [4], ChemSpider [5], and PubChem [6]; however, there are few databases that deal with polymers. MatWeb [7], managed by MatWeb LLC, contains many polymers. Focusing on engineering materials, this database has been designed for engineers, designers, and manufacturers, and currently holds information on over 175,000 materials, including nonpolymer materials such as metals, ceramics, and semiconductors. As of January 2024, 97,635 polymer materials have been registered in this database. The content is divided into several categories. For example, the most common thermoplastic is Nylon (polyamide PA) with 14,409 materials, of which more than 78% are either Nylon 66 (PA66) with 5,778 (6,984) materials or Nylon 6 (PA6) with 5,530 materials. In other words, the database contains a variety of product data on the same polymer. The units of these property data have been standardized to facilitate performance comparisons among popular engineering plastics. A catalog database is a powerful tool that allows chemical companies to develop products using industrial materials provided by stable suppliers.

An example of an industrial polymer database is CAMPUS (Computer Aided Material Preselection by Uniform Standards) [8]. Started in 1988 as a text-based information provider, this database has been improved with computer development and standardization of measurement data. However, the amount of data contained is unclear; for example, a search for Nylon 66 yielded 1,356 materials, and for Nylon 6, it yielded 1,301 materials, suggesting that the data can be overviewed from a different perspective from that of MatWeb. The data are available for download in a protected PDF format. Characteristic charts and text data are also available.

Because these databases present the properties of completed products, they do not contain much information about the manufacturing process, such as polymerization details or data on the primary structure, which is the core of polymers. In many cases, chemical or structural characteristics beyond those inferred from the product name are important for polymer chemistry and in the prediction of properties in informatics. In addition, the data are limited to well-known polymers, and may not be suitable for the development of new polymers or may require extensive extrapolation. In contrast, databases provide a significant amount of useful information for a general understanding of the properties of plastics, for application as performance benchmarks and for the development of more macroscopic industrial materials by combining materials. As discussed in the following, the data sources and policies are different from those of PoLyInfo. Nevertheless, it is important to be able to use all the available data sources in a complementary manner, and the goal should be to collaborate and not compete. Collecting diverse and voluminous data is not easy. The success of PoLyInfo is the result of a long history of curators understanding the academic significance of the comprehensive collection of polymers and jointly pioneering and sharing the advanced knowledge necessary for their classification, regardless of the data science that is currently flourishing. Continued data curation in the future will be partly supported by its inertial force, however, it will be largely driven by the fact that PoLyInfo has accompanied the humans who created plastic culture and has become a base of the lives of many people. This is more a management theory of data curation than engineering know-how. It should be noted that building a large database has management factors other than what is discussed here. In addition, PoLyInfo has a variety of tools, described in detail in section 4, to cross view the data collected in this way. This is a manifestation of PoLyInfo's philosophy that classification and structuring data, rather than

arranging them, is the most important issue. The difference between PoLyInfo and other polymer databases is that everything from data curation to publication is motivated by a consistent academic inquiry.

2. Outline of PoLyInfo

The demand for plastics has grown rapidly since the invention of synthetic polymers, which have become indispensable materials in daily life. Industrial and academic interest in polymers originates from their multiscale physical chemistry, from structural control at the molecular level to crystal structures, introduction of fillers, composite materials, and shape control, as well as the variety of functions that emerge at each structural level. National Institute for Materials Science (NIMS) has paid attention to such wide-ranging demands and has constructed a polymer database called PoLyInfo.

PoLyInfo was started in 1995 as part of the 'High Functional Basic Database' project of the Japan Science and Technology Agency (JST) and was transferred to NIMS in April 2003, which has since continued to expand the database while maintaining its own curation rules. The main features of the database are as follows:

- Fact-based curation dedicated to experimental data (excluding computational and theoretical data) described in scientific papers
- Molecular structure determination and normalization by IUPAC name as a required registration condition for structural search
- Data curation and retrieval on a per-sample basis, while maintaining bibliographic information
- Plotted data search using a unique adopted unit
- Reactant search via polymerization
- Linking to nuclear magnetic resonance (NMR) spectroscopy and infrared spectroscopy (IR) data

In particular, the normalization of all samples by the unique determination of the constitutional repetition unit (CRU) enables a systematic classification of polymers. As for the technology employed to construct the database, we achieved digitization of data collection by 2022, prepared a workflow from curation to publication in cyberspace, and are controlling progress through its monitoring. To address the recent issues in data science, we focused on data validation and governance enhancements throughout the project. This data curation system is called the PoLyInfo curation system (PoLyCun).

Figure 1 shows the workflow of PoLyCun. Each step is referred to as a 'role' in the workflow, and a different person is assigned to each role. As shown in this figure, there is a review of curation and naming, and there can be a reversion to the previous role. (1)–(3) are called 'Data curation roles' and perform the following:

- Sorting out polymers and properties for each sample
- Curating properties and related information, such as measurement conditions, from text, tables, and graphs
- For unregistered polymers, collecting the necessary information for structure identification and naming

The following (4)–(8) are called 'Lexicographic roles' and involve the following:

- Identifying the CRU of the polymers and their constitutional units (CUs) contained therein and expressing them in a registrable format, such as a simplified molecular input line entry system (SMILES)
- Verifying that the identified structures are not already registered
- Classifying polymers, copolymers, blends, and monomers

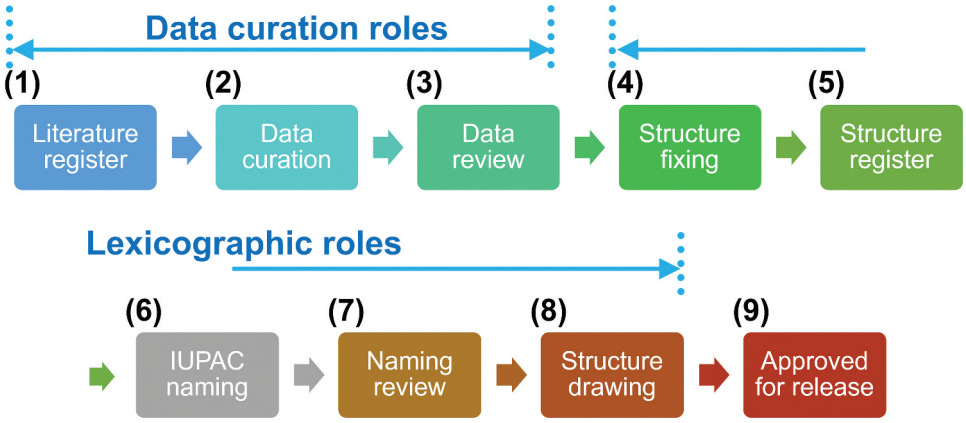


Figure 1. Workflow in the PoLyInfo curation system (PoLyCun).

- Assigning names in accordance with the International Union of Pure and Applied Chemistry (IUPAC)
- Draw a structure diagram to be displayed on the PoLyInfo graphical user interface (GUI).

The 'Data curation roles' are curated data described in the paper, whereas the 'Lexicographic roles' are originally compiled by NIMS. The first half of this workflow increases the value of PoLyInfo based on the quantity of data, and the second half further increases the value based on the quality of the data. Both these value additions were made possible through the knowledge of dedicated polymer experts.

To provide PoLyInfo with a GUI, the data systematically collected by PoLyCun were selected and compiled as appropriate and displayed in each of the screen transitions shown later. These edits on the GUI are necessary for an easy-to-understand data reference and are a manifestation of PoLyInfo's policy of not disregarding human reading, even in the age of computers handling data. The list of items displayed on the GUI based on this policy can be easily looked up in the Help of PoLyInfo (https://polymer.nims.go.jp/PoLyInfo/guide/en/help_index.html). However, when considering PoLyInfo as a subject of polymer chemistry, rather than database engineering, it would be meaningful to reorganize the data structure in terms of the data to be collected rather than the data to be displayed. Therefore, PoLyInfo (I) returns one step from providing data and reviews PoLyInfo from the PoLyCun side, that is, from the editor's viewpoint, and summarizes the information necessary for discussing polymers.

A statistical summary of PoLyInfo shows the following data counts as of January 2024.

Homopolymers: 18,697

Copolymers: 7,737

Polymer blends: 2,572

Composites: 3,069

Polymer samples: 161,464

Property points: 494,820

Literature: 20,445

The sum of the properties and polymer lexicographic data exceeds half a million, which provides an understanding of the scale of the subject addressed in this paper. A detailed breakdown of the data is provided below (https://polymer.nims.go.jp/datapoint.html).

While the details of PoLyInfo are discussed in the next sections, the accompanying NMR/IR spectral database is outlined here. PoLyInfo has long been linked to the nuclear magnetic resonance (NMR) spectra provided alongside. This provides nondestructively determined reference data on polymer stereoregularity, copolymer composition, and branching. Currently, PoLyInfo can narrow down NMR data using nuclides (^{1}H and ^{13}C) and chemical shifts specified by the region of interest (ROI). Notably, the NMR spectra presented herein differ from that collected with the cooperation of companies and are original data not collected from papers. Similarly, infrared (IR) reference spectra were released after the PoLyInfo renewal in 2022.

The number and details of NMR/IR spectra are as follows:

Total number of NMR spectra: 626

^{1}H spectra: 313

^{13}C spectra: 313

Total IR spectra: 174 (number of unique polymers: 170)

Figure 2 shows the screen transitions of the PoLyInfo GUI, where NMR/IR is described above, whereas the other items will be frequently referenced in later discussions. Before using PoLyInfo, we strongly recommend that users refer the following. Here, the database outlines and operational manuals are compiled. https://polymer.nims.go.jp/PoLyInfo/guide/en/help_index.html

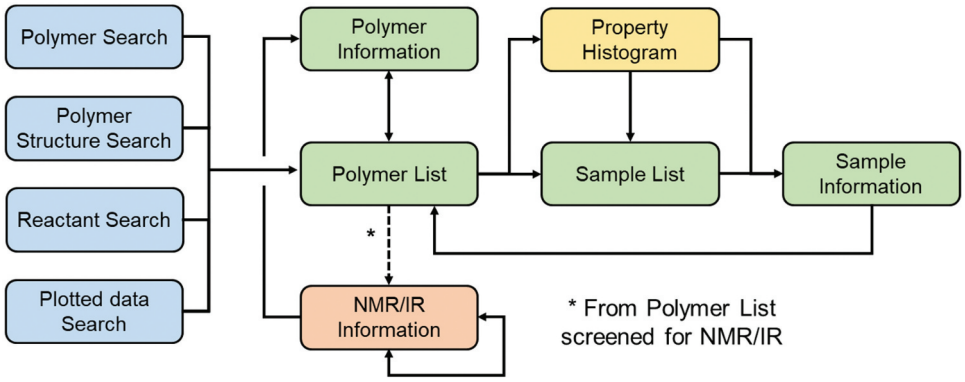


Figure 2. Screen transitions of the PoLyInfo GUI.

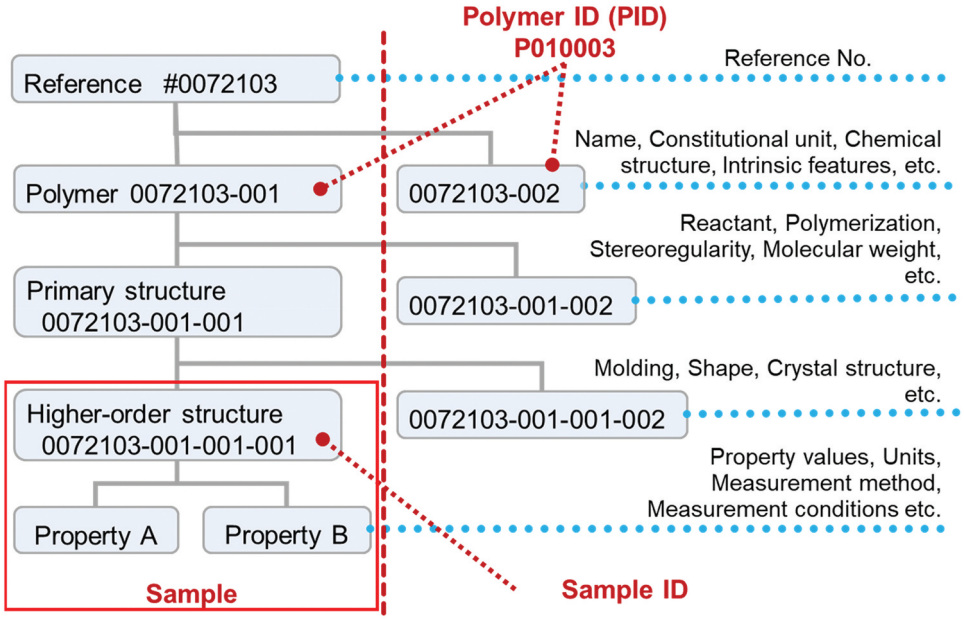


Figure 3. ID numbering rule of PoLyInfo IDs using an example: poly(but-1-ene) (P010003, 0072103-001-001-001).

3. Sample ID system

The ID system is important for understanding data management in PoLyInfo. Figure 3 shows the ID numbering rule of PoLyInfo, where the PoLyInfo sample IDs are represented by NNNNNNN-XXX-YYY-ZZZ, as in the actual representation 0072103-001-001-001.

- Because the data in PoLyInfo are curated from papers, our own literature number (NNNNNNN) is the core of the system. Sample IDs with the same literature number are described in the same paper, and PoLyInfo can identify related polymers on the GUI using this rule.
- The polymers discussed in this paper are classified and assigned with the first branch number (XXX). This step identifies polymer species with a certain CRU.
- Even if they are of the same polymer species, if the primary structure formation method, such as the polymerization method, is different, they are classified using a second branch number (YYY).
- If they have the same primary structure but different higher-order structures, a third branch number (ZZZ) is assigned.

Because most of the properties described in this paper are different for each sample determined to have a higher-order structure, the properties are considered as attributes of the sample with an ID of NNNNNNN-XXX-YYY-ZZZ. More precisely, it should be pointed out that when the same sample is measured under different conditions, it is managed internally in the database with an auxiliary number.

As an important feature of PoLyInfo, we mentioned the registration conditions for molecular structure determination and normalization by IUPAC name in Section 2. The unique molecular structure was determined using NNNNNNN-XXX, and the normalized structure was assigned a polymer ID (PID). In example 0072103–001, the IUPAC structure-based name is poly(1-ethylethylene), the IUPAC source-based name is poly(but-1-ene), and the PID is P010003. In this paper, we will use the polymer name as displayed in PoLyInfo to ensure correspondence with the PoLyInfo GUI. On the other hand, only when discussing a specific polymer, such as high-density polyethylene, we will provide the general polymer name used in the paper.

In this ID system, it is important to highlight an issue that needs to be addressed, particularly when PoLyInfo is applied in data science: In many cases, the structure determined by the CRU of the polymer, that is the PID, is used to feature the sample. However, the actual properties were provided for the sample ID, as described above. The two-level gap in Figure 3 indicates that PID alone is not sufficient as a descriptor of the properties. It is impossible to predict the actual polymer composition, particularly for industrial use, without incorporating into the descriptor the factors determining YYY and ZZZ in the sample ID.

4. Search functions in graphical user interface

PoLyInfo has the following four search functions:

(1) Polymer Search
(2) Polymer Structure Search

(3) Reactant Search
(4) Plotted data Search

These search functions output a 'Polymer list' (Figure 2), which is a list of relevant samples in PoLyInfo, grouped by a normalized polymer structure with PID as described in Section 3. The following is an overview of each search function, mainly focusing on Polymer Search and Plotted data Search, which are suitable for understanding the content of PoLyInfo.

(1) 'Polymer Search' is the default option when approaching the database from the PoLyInfo portal site. In addition to screening the recorded data by polymer name, CU, target properties, and references, it is possible to search using PID. Copolymers and blends containing the specified ID are also included in the resulting polymer list. One of the search functions narrows down the search by the number of carbon (C) atoms in the CU, which allows for a visual knowledge of the coverage of polymers handled by PoLyInfo.

Figure 4 shows the number of C atoms in CU on the horizontal axis, and the number of polymer species containing C atoms above that number (accumulated number) on the vertical axis for the homopolymers and copolymers registered in PoLyInfo. The exponential function is generally well-fitted when the number of C atoms is > 40 for homopolymers and > 10 for copolymers. Although there can be an infinite number of polymers with infinitely large CU, such polymers are impractical from either a functional or technological viewpoint. The slope of the exponential function suggests that, in general, for every 7.9 to 8.0 additional C atoms, the number of controllable polymers is halved, and this value is considered to represent the practical limit of current polymer chemistry. Extrapolating this exponential function to a C number of 1, the number of realistically controllable homopolymers was estimated to be 100,000 and 10,000 for the copolymers. Based on the difference in the regression curves, there may be a large number of unexplored homopolymers remaining in PoLyInfo's data curation capabilities. An alternative view suggests that practically promising polymers, even those for machine prediction using PoLyInfo's data, fall within this range. The difficulty of polymer informatics, even though the number is not so large for machines, is probably due to the fact that the complexity of higher-order structures is multiplied here by several orders of magnitude or more. However, for the polymer with the largest number of C atoms in the CU, the PID of the homopolymer is P462715 [9], indicating that the C number in its CU (equal to CRU for the homopolymer) C174H208N2O22 is at most 174. In terms of the copolymer P903552 [10], the C numbers in the CU components C55H32F6N4O3 (CU372694) and C118H80F12N6O9 (CU432585) are 55 and 118, respectively. Thus, the number of C atoms in the CRU is 173, and the sum of C atoms in each CU is almost the same as for the homopolymer. This C number is not far from the limit for industrially available polymers.

(2) 'Polymer Structure Search' can narrow down polymers by calculating the similarity using fingerprints with MOL files or SMILES. The well-known RDKit [11] is used here. By specifying the algorithm for generating fingerprints and a similarity metric, structures that exceed a preset threshold can be displayed. The fingerprint algorithm can be selected from 'morgan', 'daylight', 'maccs_keys',

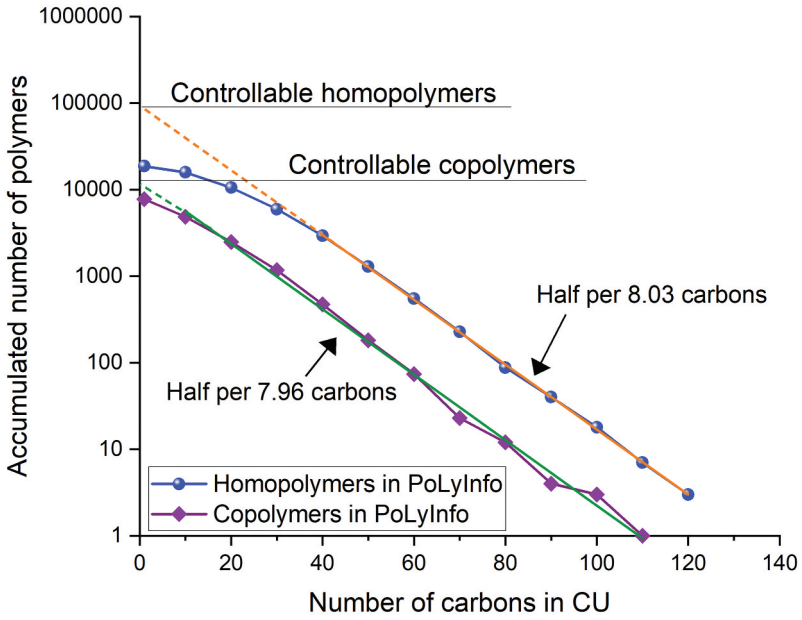


Figure 4. Exponential relationship of C in CU with the number of polymer species containing C above that number (accumulated number) for homopolymers and copolymers in PoLyInfo.

'atom_pair', 'topological_torsion', 'layered', and 'pattern', and the similarity evaluation method can be set from 'tanimoto' and 'dice'. Although the difference between the algorithms is not discussed here, the combination of 'morgan' and 'tanimoto' will be often used [12,13]. For more details, please refer to the RDKit guide on the formal site.

(3) 'Reactant Search' is realized by recording the paths from monomers to polymerized polymers. Here, the reactants to be searched can be narrowed down by 'Monomer Name', 'PubChem CID', 'Nikkaji Number', and 'Chemical Formula'. The Nikkaji Number is an ID uniquely assigned to over 3.3 million chemical substances by JST, and is linked to basic chemical information such as structural formula [14]. The idea of using well-known IDs from external databases, along with the fact that PoLyInfo does not accommodate a large number of monomers, implies an attempt to form scientific knowledge in tandem with a large external monomer database. In fact, the number of entries in PubChem at this time was 116,176,107 for compounds and 311,213,653 for substances [15]. Nikkaji is ID-linked to PubChem [16], making it understandable that the sharing of specialties is highly significant.

(4) 'Plotted data Search' is useful when simultaneously screening two target properties and qualitatively visualizing the correlation between them. Specifically, users can select any of the properties recorded in PoLyInfo (see Section 5.5) from the pull-down menu for the x- (first property) or y-axes (second property). As shown in Figure 5, plotting both the axes with the same property, i.e. the glass transition temperature (T_g), for data validation provides important insights into PoLyInfo's data curation policy. Intuitively, even though all the data should be plotted on a diagonal, much of it is found to be off-diagonal. This implies that PoLyInfo recorded multiple T_g values for each sample. In the case of composites, it is easy to understand that multiple T_g values can be observed due to material heterogeneity. However, it is necessary to note the difference from intuition in the case of neat resin. For specific verification, it is helpful to use the 'Plotted data Search' by specifying an off-diagonal area with the cursor. The results show, for

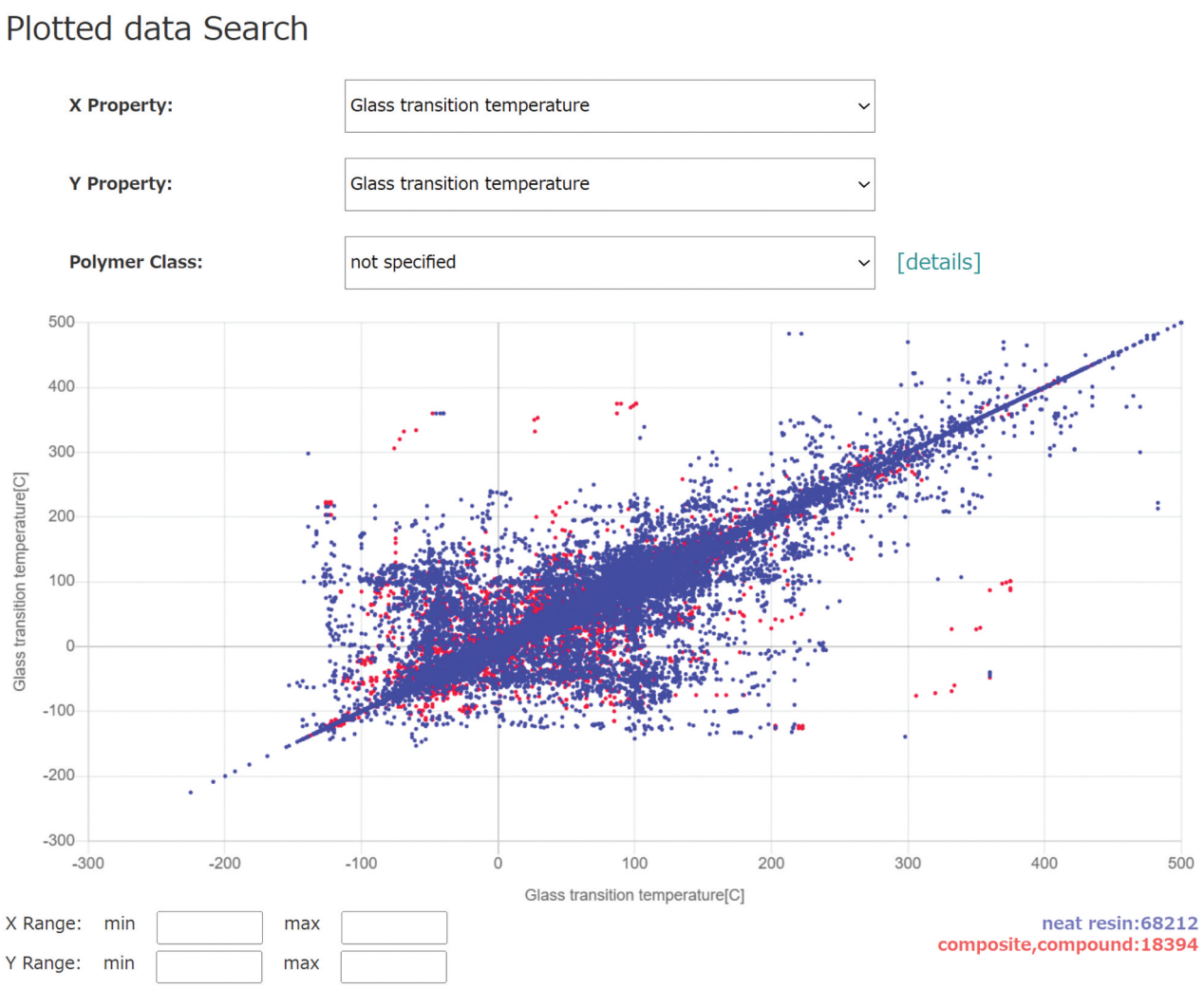


Figure 5. Screenshot of 'plotted data search' for T_g data validation.

example, for polyethene (P010001) and sample ID 0030955-001-001-002, the measured T_g values when measured using the inflection point method are –90.2°C for first cooling and 48.5°C for first heating [17]. Such multiple property values are not only recorded for T_g; there are also cases such as when density at 0% and 100% crystallinity obtained from extrapolation of the experimental results were recorded (e.g. poly(but-1-ene) P010003, sample ID 0072103-001-001-001) [18]. Consequently, even for the homopolymers of neat resins, additional descriptors may be required when PoLyInfo is used in data science. This motivated the machine-readable structuring of PoLyInfo, the details of which are discussed in part two (PoLyInfo (II)) of this series.

In addition to these search functions, PoLyInfo has a function to narrow down target samples by 'Property histogram' (Figure 2). For properties with a sufficient amount of data after the above searches, histograms can be displayed separately for the neat resin, and composite and compound. If the dispersion is reasonable, the median value can be used to approach the most normative sample data. This function can be used to determine the degree of dispersion caused by multiscale processing of polymers or by differences in the measurement conditions as discussed in the 'Plotted data Search'.

5. Database contents

5.1. Master information

PoLyInfo normalizes and masters the polymer samples. It is academically important to examine this mastering process and systematize polymers based on their basic structures. For this purpose, we further study the ID described in Section 3 and describe the PoLyInfo method employed to deal with a wider range of polymers.

The examples of IDs described in Section 3 are only for the simplest structure: homopolymer. However, actual PoLyInfo IDs include not only PIDs but also COIDs (copolymers), BDIDs (blends), and, as the basic component ID common to all of these, the constitutional unit (CUID). For the homopolymers, because the CRU is the same as the CU, it is easy to understand that the PID is simply the CUID. In fact, PoLyInfo has the same number of CUIDs as the PIDs, and PID P010003, for example, does not differ from CUID CU010003 in appearance. However, the fact that PoLyInfo explicitly separates these IDs implies that the PoLyInfo master not only classifies polymers as materials but also classifies structures at the molecular level. This becomes apparent in the case of copolymers.

In the PoLyInfo master, the CRU of the copolymers is represented by a combination of a CU and a junction unit (JU). Therefore, in the polymer information (Figure 2), which is a search result using COID, CUIDs and JUs are listed as components. In fact, PoLyInfo also assigns IDs to JUs and masters them as JUIDs; however, the GUI version of PoLyInfo does not include them in the search function, and users do not directly use them for data screening. However, it is clear that PoLyInfo's editorial policy of managing polymers at the molecular level smaller than CRUs is one of its originalities. It should be noted here that COID can be recognized as a PID whose ID starts with 'P9' and does not have a prefix like 'CO'. Furthermore, copolymer masters consider the middle-range order despite PoLyInfo's policy of molecular-level master control. For example, poly[1-carboxyethylene/1-(butoxycarbonyl)ethylene] in the IUPAC structure based name is possible to be poly[(acrylic acid)-co-(butyl acrylate)] or poly[(acrylic acid)-ran-(butyl acrylate)] in the IUPAC source based name. Although these CRUs can be represented by exactly the same combination of CU040001 and CU040006, as shown in Figure 6, they are numbered with different COIDs (the former is (a) P900001 and (b) P904994). If these middle-range orders are not specified in the paper, they are classified as 'Copolymer', but the other 'Statistical', 'Random', 'Alternating', 'Periodic', 'Block', and 'Graft Copolymer' are managed under different COIDs. Again, this demonstrates the interest in polymers as multiscale materials and the risk of using only CUIDs as structural descriptors in modeling, such as machine learning.

The BDID, which is the ID of the polymer blend, should be mentioned. For example, polyethene/polyethylene, whose BDID is BD000001, is a blend of two types of polyethene (P010001) as a component, and although they are the same, the Polymer information in PoLyInfo (Figure 2) lists P010001 twice. This indicates that the PID is only a normalized name and is

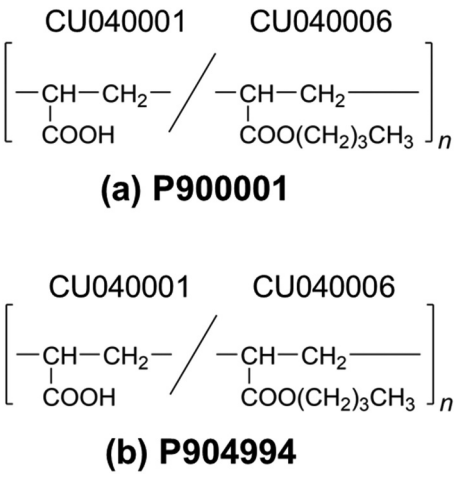


Figure 6. CRU of (a) poly[(acrylic acid)-co-(butyl acrylate)] (P900001) and (b) poly[(acrylic acid)-ran-(butyl acrylate)] (P904994) in the IUPAC source-based name.

different from the names of the actual samples. Each polymer sample was recorded in PoLyInfo with an SID even though the properties were not measured before blending. Furthermore, if the blend component is a copolymer, it can be traced back to the CUID or JUID, according to the master rules for the copolymers described above. This indicates that PoLyInfo thoroughly managed IDs at the molecular level, even for the polymer blends. For example, when searching for the ID of the repeat unit CU070013 in Polymer search (Figure 2), one can find not only the homopolymer poly(ethylene oxide) (P070013) but also the copolymer poly(ethylene oxide)-block-poly(propylene oxide)-block-poly(ethylene oxide) (P903214) that contains this CU, and poly(vinyl alcohol)//poly(ethylene oxide)-block-poly(propylene oxide)-block-poly (ethylene oxide) (BD000318), which also contains this copolymer. Figure 7 shows the hierarchy from the CU to the PID, COID, and BDID in this example. From the figure, it is clear how all the related polymers can be derived from CU070013. The master information described here can be found in 'Polymer information' in Figure 2.

5.2. Material information

Although the normalized structures and names discussed in the master information are important for polymer taxonomy, the property data in PoLyInfo are for actual samples, and the material information described herein is more detailed than that in the master information. The features of these individual samples are particularly important in the case of composites and compounds, and less so in the case of neat resins that are approximately represented by normalized features in the master information. These types are described in [Materials type].[1] In PoLyInfo, although composites and compounds have the same PID as neat resins, the effect on the properties of additives with significant industry spillover should be considered. The distribution of these material types in PoLyInfo is as follows:

Neat resin: 109,810 samples
Compound: 31,170 samples
Composite: 25,041 samples

Neat resin clearly dominates, and it can be seen that PoLyInfo has a history and policy of focusing on the academic systematization of pure polymers, rather than on the properties of materials containing polymers, which vary significantly depending on the additives used. For industrial applications, it is desirable to link to a database of composite materials such as NanoMine [19].

In PoLyInfo, the material name, amount, and shape of an additive are shown in Sample Information (Figure 2) separated by ";" [Additives]. In the database, the additives are distinguished from fillers and others and are included with related information such as their respective amounts. It is desirable to master these material names or refer to external databases.

In addition, the following items are listed as material information in PoLyInfo:

- Product name, grade, and manufacturer [Trades]
- Characteristics (general use such as 'Thermoplastic', functionality such as NLO (Nonlinear optics), crystalline state such as Liquid crystal, etc.) [Characteristics of the material]

5.3. Fabrication information

This category summarizes the information about the synthesis of each sample. The information collected by PoLyCun is as follows:

- Polymerization process information [Polymerization information]
  - Substance information regarding the polymerization process, such as the reactant name, precursor name, CU loading amount, and actual CU composition, has been described in literature.
  - Polymerization mechanism, polymerization phase, polymerization reaction type
  - Chemicals used such as catalysts, solvents, initiators, terminators, and additives, and polymerization conditions.
  - Information on dopants and doping amounts [Doping informations].
- Stereoregularity information

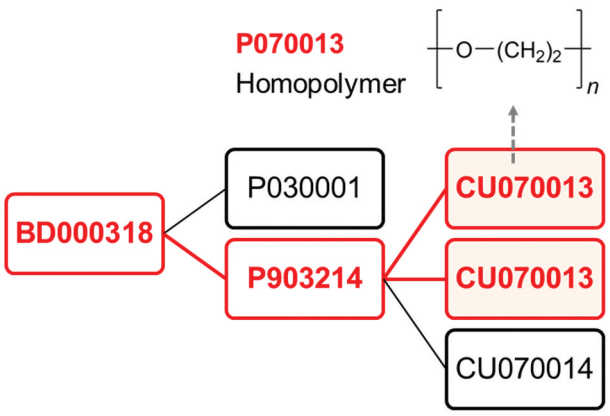


Figure 7. Structural hierarchy from CU (CU070013) to PID (P070013), COID (P903214) and BDID (BD000318) under polymer search for CU070013.

[1] In the remainder of this paper, the names of items displayed in the Sample Information in Figure 2 are listed in square brackets, as necessary, so that they correspond to the actual GUI. If there is no information on this item in the relevant sample, the item name will not be displayed on the GUI.

- Chain structure information, such as monomer sequences in copolymers, branching and cross-linking information, cross-linking density, types of end groups, names of end groups, and other information on primary structure [Primary structure information]
- Information on the average molecular weight, including classifications such as Number-average molecular weight (Mn), Weight-average molecular weight (Mw), viscosity-average molecular weight, etc., and Mw/Mn [Average molecular weights]
- Degree of polymerization information [Degree of polymerizations]
- Solution viscosity information, including classifications such as inherent viscosity, intrinsic viscosity, etc. [Solution viscosity]
- Melt flow rate measurement specifications, load, temperature, and other conditions [Melt flow rate]
- Information on the solvent [Solvent], poor solvent [Non-solvent], chemical stability, etc. [Chem. Resistance-Stable chemicals] [Chem. Resistance-Unstable chemicals]

As shown, the fabrication information is finely divided and curated in PoLyCun.

Moreover, because the average molecular weight provides important information, PoLyCun records and structures the measurement method and conditions in detail, similar to the properties information described in Section 5.5. Although the degree of polymerization, solution viscosity, and melt flow rate are curated in the same manner, they are all displayed in an unstructured list on the GUI, which is a result of prioritizing visibility over reproducing an accurate data structure. In contrast, average molecular weights other than Mw and Mn are rarely displayed because of their limited number, and it is difficult to determine their existence simply by looking through a few search results on the GUI. These facts indicate that a GUI is not necessarily the best way to understand the availability of polymer data inherent in PoLyInfo. However, implementing the GUI function to search for the details of complex fabrication information would not only require a large system but would also impair the readability and usability of the database. One service solution is to create a story that facilitates a routine search. In PoLyInfo, the samples can be narrowed by specifying a numerical range for Mw and Mn from the side menu of the Polymer List (Figure 2). This careful selection of highly important polymer information and its implementation as an additional function are also based on the service story.

To ensure the storytelling of the GUI, PoLyInfo emphasizes the terminology. For example, the types, polymerization mechanism, polymerization phase, and polymerization reaction are defined by the following links:

https://polymer.nims.go.jp/PoLyInfo/guide/jp/term_polymerization.html.

Using these items, we demonstrate the potential of PoLyInfo as a database. Table 1 is not available in the GUI but is presented as statistical information of the recorded data on the polymerization mechanism.

This list provides a quantitative understanding of the PoLyInfo scale; however, further comments should be added from the perspective of polymer chemistry. Radical polymerization has the largest number of entries, reflecting its wide industrial usage owing to its high reactivity, applicability, and resistance to polar substances such as water. However, its high reactivity makes it difficult to control stereoregularity. Considering that the stereoregularity of polymers affects various physical properties, such as the melting point, glass transition temperature, and solubility, the lack of controllability is a drawback of radical polymerization. This table summarizes the number and percentage of samples with stereospecificity information registered for each polymerization mechanism. The table shows this situation, with only 4% of the polymers of radical polymerization reporting stereoregularity. On the other hand, for coordination polymerization, in which stereoregularity can be controlled, the percentage reaches 45%. Essentially, it is desirable to be able to visualize such a polymer chemistry overview or derive knowledge from PoLyInfo, which contains a large amount of data. In PoLyInfo (II), we discuss the compensation of data-drivenness lost due to data processing in such a GUI and knowledge creation to maximize the use of the recorded data.

5.4. Formation information

In this category, information on the higher-order structures of the polymers was compiled. For synthesized polymers, mixing with other materials and molding are important for industrial applications, and the following items provide essential information for practical use:

Table 1. Correlation between polymerization mechanism and stereoregularity.

Polymerization	Number of mechanism data	Number of stereoregularity data	Percentage (%)
Radical polymerization	10554	453	4
Anionic polymerization	3042	892	29
Coordination polymerization	2830	1277	45
Living polymerization	1161	312	27
Cationic polymerization	916	120	13
Others	9243	409	4
Total	27746	3463	12

- Blending compatibility and kneading information [Mixing/Blending method]
- Molding methods such as injection and film forming, molding conditions, post-processing methods such as annealing, post-processing conditions, and sample shapes such as dumbbell and film are recorded separately. [Processing information]
- Crystal state after molding such as crystal/amorphous/liquid crystal [Crystallinities], crystal orientation [Degree of orientation]
- Classification information of crystal polymorphs such as alpha and beta [Type of crystal], crystal system [Crystal system], space groups, lattice constants, crystal interplanar spacing, measurement methods, and measurement standards [Crystallographic_informations]
- Specific rotatory power values, measurement methods, measurement standards, etc. [Specific rotatory power]
- Morphological information and related information [Morphologies]
- Other high-order structural information [High order structure crystal structures informations]

For the crystal state, orientation, crystallographic information, and specific rotatory power, PoLyCun collects measurement methods, conditions, and properties information described in Section 5.5. These are listed and displayed in the PoLyInfo GUI, with the fabrication information described in Section 5.3. This is because of the priority given to visibility; however, it should be emphasized that, from a data-driven perspective, PoLyInfo has potential that is not fully expressed by the GUI.

In terms of storyline, PoLyInfo GUI has a rough screening function for sample shape and crystallinity in the Polymer list (Figure 2). Specifically, screening using typical shapes, such as blocks, cylinders, disks, fibers, films, pellets, powders, sheets, single crystal, and solutions, is available. This indicates that information is stored independently in the database. Notably, more detailed information is stored in addition to these representative shapes, which are selected based on the usability of the GUI.

5.5. Properties information

This category summarizes the properties information handled by PoLyInfo. There are 17 major property classes in PoLyInfo, which are listed in Table 2. More than 100 properties are included in the major classes. The following is a summary of the representative properties.

Although we do not describe each of the recorded properties in this study, the details of each property (property definition, units, measurement methods, measurement standards, and essential measurement conditions) can be confirmed from the property item list in PoLyInfo.

https://polymer.nims.go.jp/PoLyInfo/guide/en/term_polymer.html#chap21

The PoLyInfo editorial policy can be found in this study. T_g, which is a typical property, is described below, and the main points are explained using the following example:

https://polymer.nims.go.jp/PoLyInfo/guide/en/property.html#P3110

Here, along with the ISO definition, the adopted unit in PoLyInfo (°C) and typical measurement methods and standards are described. Moreover, it clearly states that the property value varies significantly depending on the measurement method and test conditions, and that since the glass transition is a relaxation phenomenon, when determining T_g from dynamic measurements, it depends on the frequency. Consequently, each T_g with more than 63,000 data points has its own unique attributes (descriptors) regarding the measurement, and to accurately compare the data, descriptors unrelated to the sample

Table 2. 17 major property classes in PoLyInfo with representative properties for each class.

Property class	Representative properties
Physical property	Density
Optical property	Refractive index
Thermal property	Glass transition temperature, Melting temperature,
Electrical property	Volume resistivity, Electric conductivity, Dielectric constant
Physicochemical property	Gas permeability coefficient, Contact angle
Dilute solution property	Intrinsic viscosity, Solvent
Rheological property	Melt viscosity
Tensile property	Tensile modulus, Tensile stress (strength)
Shear property	Storage modulus
Flexural property	Flexural modulus, Flexural stress (strength)
Compression characteristic	Compressive modulus, Compressive stress (strength)
Creep characteristic	Tensile creep strain, Tensile creep compliance
Heat characteristic	Softening temperature
Impact strength	Izod impact
Hardness	Shore hardness
Heat resistance and Combustion	Oxygen index
Other property	Half-value dose, PVT relation

structure must be considered. These descriptors must be extracted from the Sample Information (Figure 2). Furthermore, there are several physical properties with external factors that require modeling, as categorized below:

(1) Properties expressed as model parameters
The crystallization kinetics is expressed by the parameter set of Avrami equation.
The gas diffusion coefficient is expressed as a characteristic quantity in the Arrhenius plot.
(2) Dynamic properties
Properties resulting from various relaxation processes in the electrical, rheological, and mechanical fields include dielectric dispersion, dynamic viscosity, and various dynamic mechanical properties.
(3) Properties dependent on other materials
Intrinsic viscosity, which depends on the solvent.
Gas permeability coefficient, etc., the physical properties of which vary depending on the permeating gas species.
(4) Physical properties defined by other processes
The G value representing molecular changes induced by radiation.
(5) Properties defined by others
The thermal decomposition temperature records the weight loss due to temperature, along with the measurement temperature.

Finally, we mention properties that are not systematically collected in PoLyInfo. PoLyInfo actively provides such properties that are highly correlated, have a significant impact on the recorded properties, or promote the understanding of polymers as 'Related Information' at the end of Sample Information. In particular, as mentioned in Section 2, PoLyInfo records properties at the sample level, while at the same time managing samples at the literature level. This means that the correlations between the physical properties discussed in this paper can be summarized on the GUI. For example, for high-density polyethylene (PID: P010001, Sample ID: 0014069-001-001-001 ~ 004), the effects on crystallinity, density, and thermal and mechanical properties for various sunlight exposure periods [20] are summarized in a table to show the weather resistance. Although it would be difficult to create a large database for weather resistance, it should be emphasized that PoLyInfo reproduces useful information systematically summarized in a single paper on a GUI.

6. Challenges and prospects

With the progress of polymer chemistry, polymers are exhibiting increasingly diverse structures. Ladder polymers are expected to be developed as low-dimensional conductive materials owing to their multiplicity and symmetry [21], and star polymers are also expected to develop new properties owing to their restricted segmental motion [22]. PoLyInfo also contains nonlinear polymers. As mentioned previously, PoLyInfo determines the structures of all the registered polymers. However, due to the increasing complexity of these structures, there are cases where the IUPAC nomenclature [23] cannot be applied.

As shown in Figures 8a & 8b for ladder-poly{1,1'-dimethyl-1,1'-[(2-tert-butylanthracene-9,10-diyl)di(4,1-phenylene)]di(silanediol)} (P472330) [24] and ladder-poly{N,N'-bis[dichloro(phenyl)silyl]benzene-1,4-diamine} (P472332) [25], although the ladder structure is represented by 'ladder-' prefix in the IUPAC source-based name, an appropriate structure-based name has not been determined because of the benzene ring in the bridge chain. The structure-based name has also not been determined for the copolymer 4-star-poly(hxano-6-lactam)-nu-alpha-{2-[3-(3-isocyanato-4-methylphenyl)-1-(3-isocyanato-4-methylphenylcarbamoyl)ureylene]propyl}-omega-[3-(3-isocyanato-4-methylphenyl)-1-(3-isocyanato-4-methylphenylcarbamoyl)ureylene]poly(oxyethylene) (P907957) [26], which has the star structure, as shown in Figure 8c. The occasional revision of the IUPAC recommendations [27] also raises questions regarding the manageability and permanence of polymeric names in the databases. Eventually, structural representation by SMILES or MOL may remain universally

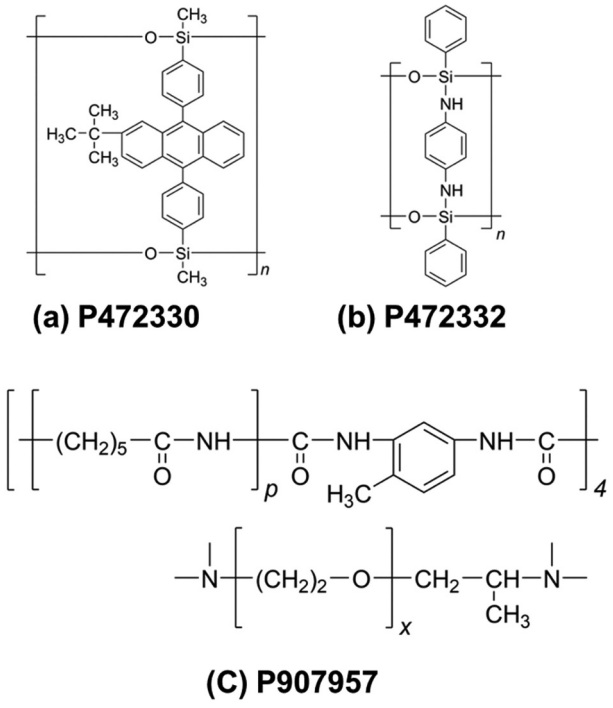


Figure 8. CRU examples of a polymer without a definable IUPAC structure-based name; CRUs for ladder polymer (a) P472330 and (b) P472332, and CRU for star polymer (c) P907957.

relevant. However, even these hashed molecular structures do not fully represent all the polymers. For example, structures, such as poly(vinyl ferrocene) (P039034) [28] and poly[hydroquinone-alt-(terephthaloyl dichloride)] (P099041) [29], as shown in Figures 9a & 9b, where chemical bonds are not definite, introduce a principle deficit in the coverage of structure determination, which is an advantage of PoLyInfo. Even though SMILES Arbitrary Target Specification (SMARTS) [30], an extension of SMILES, is excellent for the purpose of search because it can handle multiple structures, it is not sufficient for the purpose of comprehensive structure management of polymers and notation in a GUI. Currently, no method has been found to describe all structures in PoLyInfo.

While PoLyInfo's meticulous data collection has deepened our understanding of polymers, it has also fallen into the endless task inherent in polymers of picking one out of an infinite number of combinations of polymers by chance. We also feel that algorithms that guarantee the uniqueness of molecular structures are reaching their limits. Although an attempt to collect all the complex structures may be the ultimate goal in taxonomy, we are convinced that our goal is to obtain the general truth from the many examples we have accumulated so far; that is, to create polymer science knowledge as well as to continue data curation that will keep pace with the future development of polymer chemistry.

7. Conclusion

This article provides an overview of PoLyInfo, which is a polymer database of National Institute for Materials Science, containing more than half a million data points, from the viewpoint of a data editor. It describes how polymers, for which multiscale physics and chemistry produce various functions, are edited in the database and how they are provided. While visibility and storytelling are emphasized in a GUI, the collection of data is managed by PoLyCun, an accurate workflow management system in cyberspace. A considerable amount of research work has been done not only on the mere collection of polymer data described in literature but also on the systematization of the structure and unique naming of polymers to create lexicons, resulting in the originality of PoLyInfo.

However, information on raw monomers is more abundant in external databases, and database linkages that balance originality and collaboration are expected to generate significant chemical knowledge. In our subsequent paper PoLyInfo (II), we present a machine-readable attempt to achieve this feat.

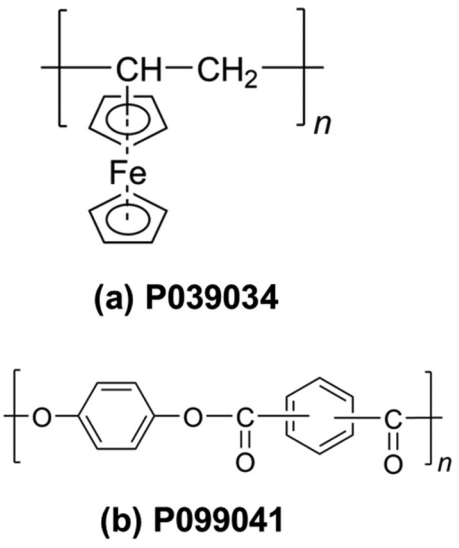


Figure 9. CRU examples of a polymer without definable SMILES; CRUs for (a) P039034 including ferrocene and (b) P099041 including phthalate.

Acknowledgements

The polymer database PoLyInfo is managed and operated by National Institute for Materials Science (NIMS) and is promoted as a project of NIMS using its own financial resources.

Disclosure statement

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

ORCID

Masashi Ishii http://orcid.org/0000-0003-0357-2832
Isao Kuwajima http://orcid.org/0000-0002-5994-3834

References

[1] PoLyInfo [Internet]. Tsukuba, Japan: NIMS; [cited 2024 Feb 14]. Available from: https://polymer.nims.go.jp/
[2] Nakamura Y, Gros A, Zhang W, et al. Exception search in databases for polymers with practically contradictory properties of heat resistance and transparency. Polym Chem. 2023;14(33):3881–3887. doi: 10.1039/D3PY00565H
[3] Yuan W, Hibi Y, Tamura R, et al. Revealing factors influencing polymer degradation with rank-based machine learning. Patterns. 2023;4(12):100846. doi: 10.1016/j.patter.2023.100846
[4] J SciFinder [Internet]. Columbus (OH): American Chemical Society; [cited 2024 Feb 14]. Available from: https://scifinder-n.cas.org
[5] ChemSpider [Internet]. London, UK: Royal Society of Chemistry; [cited 2024 Feb 14]. Available from: http://www.chemspider.com/
[6] Kim S, Chen J, Cheng T, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373–D1380. doi: 10.1093/nar/gkac956
[7] MatWeb 'Online Materials Information Resource - MatWeb' [Internet]. Blacksburg (VA): MatWeb,

LLC; [cited 2024 Feb 14]. Available from: https://www.matweb.com/index.aspx/
[8] CAMPUSplastics [Internet]. Frankfurt am Main, Germany: CWF GmbH; [cited 2024 Feb 14]. Available from: https://www.campusplastics.com/
[9] Goto H, Akagi K. Synthesis of a pyrrole-based methine bridge type liquid-crystalline conjugated polymer. J Polym Sci A Polym Chem. 2004;43 (3):616–629. doi: 10.1002/pola.20574
[10] Thelakkat M, Pösch P, Schmidt HW. Synthesis and characterization of highly fluorescent main-chain copolyimides containing perylene and uiqnoxaline units. Macromolecules. 2001;34(21):7441–7447. doi: 10.1021/ma010615w
[11] RDKit: Open-source cheminformatics [Internet]. San Francisco (CA): GitHub, Inc.; [cited 2024 Feb 14]. Available from: http://www.rdkit.org
[12] Morgan HL. The generation of a unique machine description of chemical structures — a technique developed at chemical abstracts service. J Chem Doc. 1965;5(2):107–113. doi: 10.1021/c160017a018
[13] Willett P. Similarity-based virtual screening using 2D fingerprints. Drug Discov Today. 2006;11(23–24):1046–1053. doi: 10.1016/j.drudis.2006.10.005
[14] Nikkaji web [Internet]. Tokyo, Japan: Japan Science and Technology Agency (JST); [cited 2024 Apr 25]. doi: 10.15079/NIKKAJI
[15] PubChem Data Counts [Internet]. Bethesda (MD): National Library of Medicine; [cited 2024 Feb 14]. Available from: https://pubchem.ncbi.nlm.nih.gov/docs/statistics
[16] NBDC Nikkaji RDF, Links to Other DBs (based on PubChem) [Internet]. Tokyo, Japan: Department of NBDC Program, Japan Science and Technology Agency (JST); [cited 2024 Feb 14]. Available from: https://dbarchive.biosciencedbc.jp/en/nikkaji/data-7.html
[17] Danch A, Osoba W. Thermal analysis and free volume study of polymeric supermolecular structures. J Therm Anal Calorim. 2004;78 (3):923–932. doi: 10.1007/s10973-005-0458-0
[18] Nishioka A, Yanagisawa K. Crystallinity and stereoregularity of polybutene-1 I. Crystallinity. Kobunshi Kagaku. 1962;19(211):667–675. doi: 10.1295/koron1944.19.667
[19] Zhao H, Wang Y, Lin A, et al. NanoMine schema: an extensible data representation for polymer nanocomposites. APL Mater. 2018;6(11):111108. doi: 10.1063/1.5046839
[20] Jabarin SA, Lofgren EA. Photooxidative effects on properties and structure of high-density polyethylene. J Appl Polym Sci. 1994;53(4):411–423. doi: 10.1002/app.1994.070530404
[21] Del Valle MA, Silva ET, Diaz FR, et al. Electropolymerization of 2,3-diaminophenol. J Polym Sci A Polym Chem. 2000;38(9):1698–1703. doi: 10.1002/(SICI)1099-0518(20000501)38:9<1698:AID-POLA36>3.0.CO;2-3
[22] Qian Z, Minnikanti VS, Sauer BB, et al. Surface tension of symmetric star polymer melts. Macromolecules. 2008;41(13):5007–5013. doi: 10.1021/ma8002888
[23] Favre HA, Powell WH. IUPAC recommendations and preferred names 2013. Cambridge (UK): RSC Publishing; 2014.
[24] Zhang J, Zhize C, Wenxin F, et al. Supramolecular template-directed synthesis of stable and high-efficiency photoluminescence 9,10-diphenylanthryl-bridged ladder polysiloxane. J Polym Sci A Polym Chem. 2010;48 (11):2491–2497. doi: 10.1002/pola.24021
[25] Zhang T, Deng K, Zhang P, et al. Supramolecular template-directed synthesis of perfect phenelenediimino-bridged ladderlike polyphenylsiloxanes. Chem Eur J. 2006;12 (13):3630–3635. doi: 10.1002/chem.200501447
[26] Xiang M, Xu S, Li C. Monomer casting nylon-6-b-polyether amine copolymers: synthesis and antistatic property. Polym Eng Sci. 2016;56 (7):817–828. doi: 10.1002/pen.24310
[27] IUPAC: Recommendations and Technical Reports [Internet]. Research Triangle Park (NC): International Union of Pure and Applied Chemistry; [cited 2024 Feb 15]. Available from: https://iupac.org/what-we-do/recommendations/
[28] Pittman CU Jr., Lai JC, Vanderpool DP, et al. Polymerization of ferrocenylmethyl acrylate and ferrocenylmethyl methacrylate. Characterization of their polymers and their polymeric ferricinium salts. Extension to poly(ferrocenylethylene). Macromolecules. 1970;3(6):746–754. doi: 10.1021/ma60018a007
[29] Gerhard EFL, Kurt FR. Characterization of novel polymers by softening behavior, thermogravimetric analysis, and differential thermal analysis. J Appl Polym Appl Polym Symp. 1969;8:171–188.
[30] SMILES Arbitrary Target Specification (SMARTS) [Internet]. Laguna Niguel (CA): Daylight Chemical Information Systems, Inc.; [cited 2024 Apr 25]. Available from: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html