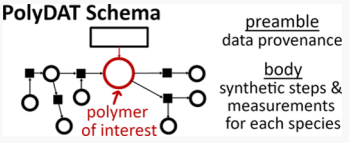
# PolyDAT: A Generic Data Schema for Polymer Characterization

Tzyy-Shyang Lin[1,2], Nathan J. Rebello[1,2], Haley K. Beech[1,2], Zi Wang[1,3], Bassil El-Zaatari[1,4], David J. Lundberg[1,2,] Jeremiah A. Johnson[1,4], Julia A. Kalow[1,4], Stephen L. Craig[1,3], and Bradley D. Olsen[1,2]*

[1]NSF Center for the Chemistry of Molecularly Optimized Networks; [2]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; [3]Department of Chemistry, Duke University, Durham, North Carolina 27708, United States; [4]Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States; [5]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

**ABSTRACT:** Polymers are stochastic materials that represent distributions of different molecules. In general, to quantify the distribution, polymer researchers rely on a series of chemical characterizations that each reveal partial information on the distribution. However, in practice, the exact set of characterizations that are carried out, as well as how the characterization data are aggregated and reported, is largely nonstandard across the polymer community. This scenario makes polymer characterization data highly disparate, thereby significantly slowing down the development of polymer informatics. In this work, a proposal on how structural characterization data can be organized is presented. To ensure that the system can apply universally across the entire polymer community, the proposed schema, PolyDAT, is designed to embody a minimal congruent set of vocabulary that is common across different domains. Unlike most chemical schemas, where only data pertinent to the species of interest are included, PolyDAT deploys a multi-species reaction network construct, in which every characterization on relevant species is collected to provide the most comprehensive profile on the polymer species of interest. Instead of maintaining a comprehensive list of available characterization techniques, PolyDAT provides a handful of generic templates, which align closely with experimental conventions and cover most types of common characterization techniques. This allows flexibility for the development and inclusion of new measurement methods. By providing a standard format to digitalize data, PolyDAT serves not only as an extension to BigSMILES that provides the necessary quantitative information but also as a standard channel for researchers to share polymer characterization data.




## 1. INTRODUCTION

Having accessible data that is well-structured is the foundation for cheminformatics. For most fields of chemistry, structured datasets can be readily curated by collecting data using a molecule-property tuple/pair style format that relates desired properties with the structures of the molecules of interest. Data in this format fits naturally into widely available and well-supported relational database technologies, which store data in series of data tables relating molecular properties with the corresponding chemical descriptors.[1] Assimilating data from different sources is straightforward as the chemical descriptors can be used unambiguously to define the chemical system and provide a handle to collate and aggregate distinct instances of data for the same chemical object. In practice, this system relies on the existence of a chemical representation that encodes the unique chemical structure of the species of interest. For most molecules, this information is conveniently encoded with representations that detail their chemical connectivity, such as SMILES (simplified molecular-input line-entry system) strings[2,3] for organic molecules, nucleic acid sequences for RNA, or amino acid sequences for proteins.[4−7]

However, this paradigm for collating data according to molecular descriptors cannot be translated to polymers. This challenge comes from the fact that polymers are ensembles of molecules produced by stochastic systems of reactions, meaning that there is no single representation that can capture the full molecular detail of a polymer. For instance, sequence representations are rarely applicable for polymers as polymer chains do not have deterministic sequences and lengths. Similarly, line notations such as SMILES do not apply to polymers as they describe deterministic connectivity. In practice, unless it is an aliquot of another polymer, each polymer can be effectively regarded as a unique chemical object. This issue makes polymer data largely disparate and presents significant challenges to curating high-quality polymer datasets. Both the magnitude of the challenge and of the associated lost opportunities are growing as advances in polymer design and synthesis enable an ever-increasing range of chemical structure and function to be incorporated into polymers. The development of a good solution that would address these complexities is therefore paramount to the progress of polymer informatics.

Currently, for most available datasets, polymer entries are identified through names of the polymers.[8−14] Name-based identification, however, often leads to ambiguity in molecular structure specification[15] because polymer chemistry poorly adheres to IUPAC polymer nomenclature,[16,17] making automatic translation between polymer names and structure difficult. Furthermore, in practice, polymer names often do not

have sufficient descriptive power, and multiple molecular ensembles often correspond to a single name. While the recently developed BigSMILES line notation,[18,19] which extended the syntax of SMILES and included random graph operators to capture the stochastic nature of polymers, addressed several shortcoming of name-based descriptors, the many-polymer-to-one-descriptor issue still persists. In general, a BigSMILES string only specifies the set of possible molecules that constitute a polymer, therefore providing only a qualitative description of the molecular ensemble. To characterize a polymer fully, the probability and weight must be assigned to each of the molecule within the molecular ensemble associated with the polymer.

Because of the wide chemical and structural diversity presented in polymeric systems and the stochastic nature of the reaction products, complete quantification of the molecular ensemble is generally difficult and expensive, if not impossible with existing techniques. Instead of trying to completely characterize the ensemble (i.e., providing a probability of occurrence for each molecular structure within the set), polymer researchers usually describe the structures of polymers by enumerating multiple characterizations that each provide partial knowledge on the probability distribution of the underlying ensemble.[20,21] For example, many measurements of molar mass report only moments of the distribution,[22] and even measurements of the full distribution may not provide information on specific monomer arrangements or compositional variations.[23] Moreover, it is also common for characterization to be performed indirectly: instead of directly characterizing the polymer of interest, measurements are performed on precursors[24−26] or post-functionalized counterparts from which information for the polymer of interest may be inferred.[25] The diversity of materials and applications has therefore led to a great deal of divergence in which characterization is performed and how data is reported.[11,27,28] In general, characterization data from different sources is not similarly organized, and different sources usually provide different sets of characterization. This lack of a shared schema introduces significant obstacles to the dissemination of data and hinders the establishment of a common data repository. Moreover, most literature data are currently reported in a written format that is intended for human readers, and converting these data into a computer-friendly format is generally a difficult task. This greatly restricts the use of computer-based tools and renders almost all advanced analytics tools inapplicable. These limitations largely prevent the adoption of data-driven research tools in polymer science. Therefore, developing a common digital schema that addresses the issue of disparate data and other challenges associated with the curation of polymer data is critical to the polymer community.

To lower the barrier for data dissemination and promote polymer informatics, a generic data schema that serves as a universal polymer system of record is proposed in this work. Unlike existing chemical schemas,[27,29−31] which mostly focus on reporting details pertinent to the species of interest, the proposed schema, PolyDAT simultaneously reports the characterizations of multiple species relevant to the polymer of interest as well as reactions and processing procedures that provide important relational information between these species. Notably, while the Chemical Markup Language (CML),[32,33] in association with its modules CMLReact[34] and PML (Polymer Markup Language),[35] can also be used as a standard way for reporting the synthetic procedure for a polymer, the design philosophy and implementation is appreciably different between PML and PolyDAT. In general, PML is more concerned about providing a set of computable functions that could be used to generate realizations of a polymer. As such, the central parameters of concern to PML are the transitional probabilities that provide a recipe for computationally generating a polymer chain. In contrast, PolyDAT is mainly concerned about experimental data, and the fundamental philosophy for PolyDAT is to provide a schema that is closely aligned with experimental practices. Therefore, in PolyDAT, an experiment-centric approach is taken to record the characterization for each species. This experiment-centric approach has been found highly useful for capturing the contextual information about materials and chemicals. In this realm, the Graphical Expression of Materials Data (GEMD)[36] schema has seen much success in providing a system of record for industrial materials. However, while GEMD provides comprehensive support for capturing the synthetic and processing history of materials, its support for the characterization of structure and composition is relatively basic. Since such information is paramount to many polymer applications, in PolyDAT, fields pertinent to the characterization of chemical structure and composition are explicit considered. Overall, PolyDAT complements existing schema such as GEMD and PML by providing direct support for the experimental characterization of the structure of polymer materials.

Within a PolyDAT document, structural information for each polymer is encoded through a series of experimental measurements in parallel. This format guarantees flexibility for users to incorporate any combination of characterization data, including data from future characterization techniques that have yet to be developed. Furthermore, the schema is designed so that most common characterizations are collapsed into a handful of categories and easily encoded with a set of standardized data structures. This design enables the comparison across a wide range of polymers characterized with distinct characterization techniques that would have been otherwise difficult. Moreover, by offering a standard data encoding scheme, the schema provides the necessary infrastructure for building computer-based tools. Finally, this project represents an initial endeavor toward identifying the commonalities across the highly nonhomogeneous and disparate polymer literature. In the long term, PolyDAT and subsequent efforts will constitute a set of congruent languages for all domains of the polymer community, thereby improving integration and allowing better comparison across parallel studies. Overall, these efforts are motivated by the pursuit of a FAIR (findable, accessible, interoperable, and reusable)[37,38] data model for polymeric data, which is vital to the assimilation of polymer data sets that can be very useful in applications such as the high-throughput screening of biomaterials,[39] automated discovery of novel materials for energy applications,[40] or molecular design of high-performance polymer membranes.[41]

## 2. METHODS

**2.1. Schema Structure.** A PolyDAT object is a data repository that encapsulates all relevant chemical characterization data for a polymer sample and relevant precursors or post-modified species as well as the chemical relationships between the sample of interest and other relevant species. To

accomplish this, PolyDAT records data in a hierarchical structure (Figure 1). The outermost layer of PolyDAT consists

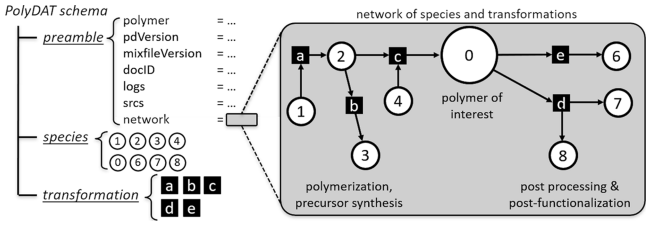



**Figure 1.** Illustration of the data structure of the PolyDAT schema. The overall object, which serves as a data repository, consists of a preamble, a species section, and a transformation section. The preamble is composed of six fields: *polymer*, *pdVersion*, *mixfileVersion*, *docID*, *logs*, *srcs*, and *network*. The first six elements provide critical metadata, and the *network* field provides a summary of how different species found within the species section are related to each other through transformations such as chemical reactions or other physicochemical processes such as separations.

of three major components. First, a *preamble* section provides data provenance, metadata, and other information essential to the parsing of the document. Second is a *species* section where characterization for individual polymer species is provided. Finally comes a *transformation* section where details on distinct chemical reactions and other relevant physicochemical processes are provided. These transformations specify how multiple species provided within the species section are related to each other. A sample document encoded with PolyDAT is provided in Figure 2 in JavaScript object notation (JSON) format.[42] Throughout this work, the JSON format will be used to illustrate the application of PolyDAT. However, it should be noted that the proposed schema provides a generic data structure specifying how data could be organized, and it could be implemented using other file formats such as the extensible markup language (XML)[43] or YAML[44] as well.

**2.2. Preamble Section.** The preamble section is designed to encapsulate metadata that provide critical data provenance to the overall PolyDAT object and to provide a summary and organization for the data found within the other sections. Within the preamble section, the seven essential entries in the following pattern should always be provided:

```
preamble = {
    "polymer" : string,
    "pdVersion" : number,
    "mixfileVersion" : number,
    "docID" : string,
    "logs" : array of log-obj,
    "srcs" : array of src-obj,
    "network" : array of strings
}
```

The *polymer* entry, which contains the BigSMILES representation for the polymer of interest, provides a basis for finding the document through structural search. BigSMILES is an extension of SMILES. For an illustration of the basic syntax of BigSMILES, please refer to Figure 3. The *pdVersion* entry specifies the version of the PolyDAT schema used by the document. While future revisions of the schema should mostly be backward compatible, this field provides delineation in case syntactical ambiguity occurs. The *mixfileVersion* field specifies the version of Mixfile format[31] utilized within the *species* section; details for the use of Mixfile are provided in the next subsection. The *docID* field consists of a document identifier string. Analogous to the protein ID in the Protein Data Bank, this field provides a unique identifier for unambiguous reference to the document. The *logs* entry consists of an array of log objects (*log-obj*), providing a detailed record for each revision to the document. Good log design should include a clear revision history as well as unique identifiers to individuals or organizations responsible for the provenance of the data (i.e., an ORCID ID). Similarly, the *srcs* field, which consists of an array of source objects (*src-obj*) provides a container for explicitly keeping track of the individual sources from which the data compiled was extracted. Examples of the log object as well as the source object are illustrated in Figure 2. Details on the format of the log object and the source object are provided in the Supporting Information.

Finally, in the *network* field, the species and transformations found within the corresponding sections are declared as an array of transformations. As illustrated in Figure 2, each transformation is encoded as a Reaction SMILES-like[46] string. Each string is composed of two dot-delimited lists of species, corresponding to the reactants and the products, respectively, separated by the transformation enclosed between two right angle brackets. For brevity, each species or transformation is associated with a unique identifier. These placeholder identifiers can be any alphanumeric strings encapsulated within a pair of square brackets, except for the polymer species of interest, which should always be denoted as "[0]". Note that the *network* array can be empty in the case that the reaction network is consisted of only the polymer of interest and no transformation. While only seven essential entries are included, the preamble is easily expanded to include additional entries. Data such as IUPAC or trade names for polymers or additional text description of the polymer could also be incorporated into the preamble section as desired. An example illustrating the usage of the preamble section is provided in Figure 2a.

**2.3. Species Section.** The species section provides detailed characterization for individual species declared within the *network* section of the preamble. The species entry is composed of an array of species objects (*species-obj*). Each species object within the array has the following pattern:

```
species-obj = {
    "ID" : string,
    "contents" : array of component-obj
}
```

where *ID* is the placeholder identifier string matching that used within the preamble and *contents* is an array of component objects (*component-obj*), with each array element corresponding to a single component within the species. An example on using the provided syntax to describe a pure poly(ethylene glycol) species is shown in Figure 2b. Note that this template is meant to be a minimal construction, and individual users are allowed to provide additional fields that are suitable for their specific applications. For instance, entries such as catalog, lot, or batch numbers that provide additional information can be incorporated as well.

Apart from describing single component systems, this multi-component design also provides general support for mixtures. An example for specifying a mixture is provided in Figure 2b. As illustrated in the second species, species [1], within the example, more than one component can be specified, along with their concentration, by providing multiple entries under the *contents* entry. This design allows issues of purity and transformations by separation to be captured in the schema. It

a)

```
{
   "preamble" : {
      "polymer" : "[H]{[>][<]NCC(=O)[>][<]}O",
      "pdVersion" : 1.0,
      "mxfileVersion" : 0.01,
      "docID" : "doc-xxx.xxx.xxx",
      "log" : [   {  "author" : [ "ORCID:https://orcid.org/0000-0002-7272-7140" ],
                     "date" : "2020-03-20",
                     "msg" : "document first created" }   ],
      "srcs": [   {   "citeID": "olsen2019", "doi": "https://doi.org/10.1021/acscentsci.9b00476",
                     "desc": "Specification of BigSMILES Syntax"   }   ],
      "network" : [ "[1]>[a]>[2]", "[2]>[b]>[3]", "[2].[4]>[c]>[0]"], "[0]>[e]>[6]", "[0]>[d]>[7].[8]" ]
   },
   "species" : [ ... ],
   "transformation" : [ ... ]
}
```

b)

```
{
   "preamble" : { ... },
   "species" : [
      {   "ID" : "[0]", "contents" : [
            {   "ID" : "[0:1]", "bigsmiles" : "O{[>][<]CCO[>][<]}[H]"    }
         ]
      },
      {  "ID" : "[1]", "contents" : [
            {    "ID" : "[1:1]", "bigsmiles" : "O{[>][<]CCO[>][<]}[H]",
                 "quantity" : 60, "units" : "UO_0000076" },
            {    "ID" : "[1:2]", "bigsmiles" : "O{[>][<]CC(C)O[>][<]}[H]",
                 "quantity" : 40, "units" : "UO_0000076" }
         ]
      }
   ],
   "transformation" : [ ... ]
}
```

**Figure 2.** Illustrative examples of the usage of the PolyDAT preamble and species sections. Chemical structures for corresponding examples are provided in shaded boxes besides the BigSMILES strings for readers' reference. (a) Example preamble for a poly(glycine) sample, in which the network section corresponds to the reaction network illustrated in Figure 1. (b) Example species section with two species. The first species, species [0], corresponds to a single component poly(ethylene glycol), whereas the second species, species [1], corresponds to a two component 60/40 mole fraction blend of poly(ethylene glycol) and poly(propylene glycol). Note that the units for the blends are given in the universal resource identifiers (URIs) of the corresponding object in the Unit Ontology (UO).[45] Here, UO_0000076 corresponds to mole fraction. The corresponding URIs for common units are given in the Supporting Information.

should be noted that individual components need not be pure substances. As illustrated by the *component-obj* pattern below, a component can either be a substance that can be specified by a single molecular identifier, e.g., a SMILES string representing a small molecule or a BigSMILES string for a polymer, or alternatively be a finite-component mixture whose contents are determined by an array of constituent sub-components. Note that in the former case, species may correspond to pure substances or polymeric molecular ensembles that are not discrete mixtures. These non-pure components can be arbitrarily nested within each other, allowing the definition of mixtures in a hierarchical manner. Within each component object, apart from the molecular structural specification of the chemical identity, each object also encapsulates a series of characterization data that provide further insights into the chemical structure of the component. Overall, the general pattern for *component-obj* is defined by the following sub-schema:

```
component-obj = {
        "ID" : string,
        "bigsmiles" : string,
        "name" : string,
        "description" : string,
        "quantity" : number(s),
        "units" : string,
        "contents" : array of component-obj,
        "empirical_formula" : string,
        "characterization" : {
           "Mw/Mn" : array of scalar-obj,
           "Mn" : array of scalar-obj,
           "Mw" : array of scalar-obj,
           "Mz" : array of scalar-obj,
           "DPn" : array of scalar-obj,
           "DPw" : array of scalar-obj,
           "DPz" : array of scalar-obj,
           "skewness" : array of scalar-obj,
           "kurtosis" : array of scalar-obj,
           "MWD" : array of vector-obj,
           "ratios" : array of ratio-obj
        }
    }
```

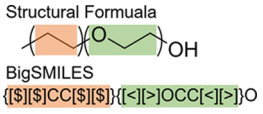



**Figure 3.** Illustrative example for the syntax of BigSMILES. In BigSMILES, curly brackets are used to denote stochastic objects. Each stochastic object represents a polymeric segment with indefinite number of repeat units. Within a stochastic object, a list of repeat units is provided. For the example presented in this figure, as there are two polymeric blocks, the BigSMILES is composed of two stochastic objects. Within the left block, one repeat unit, [$]CC[$], is specified. Similarly, [>]OCC[<] is specified for the other block. [$], [>], and [<] represent bonding descriptors that specify the permissible connectivity patterns between repeat units. [$] represents sites that can be joined with any other [$] sites, whereas [>] is only allowed to connect with [<], and vice versa. The bonding descriptors at the terminals of the stochastic objects denote how the connectivity pattern between repeat units end groups. For instance, the rightmost [>] denotes that the OH end group must be connected to an atom that is tethered to the [<] descriptor. In this case, it can only connect with the carbon end. For more details on BigSMILES, please refer to refs 18 and 19.

Within *component-obj*, each component is given a unique identifier under the *ID* field. To ensure identifiers are unique within a PolyDAT file, they should be assigned in the following convention: the ID of a (sub-)component should retain the ID of its parent (sub-)component or parent species, less the terminal brackets, followed by a colon and a (sub-)component specific alphanumeric string. For instance, every component within species with ID "[1]" will have the form "[1:*compX*]", whereas the sub-components of a component with ID "[1:comp1]" will take the form "[1:comp1:*subcompY*]", where *compX* and *subcompY* are alphanumeric strings unique across all components or sub-components associated with the same parent. For a pure component or a polymer component, the chemical identity of the species is specified by providing the chemical structure, in BigSMILES format, within the *bigsmiles* field. To provide a basis for specifying atom mapping in the transformation section, atoms within the BigSMILES strings can be labeled with alphanumeric identifiers. The atom labels are arbitrary given that they are unique across the component or sub-component. In the case where the material or chemical entity is not characterized by a chemical structure, the *empirical_formula* field can be used in place of *bigsmiles* to specify composition of the material. Furthermore, the *name* field can also be used in place of or in conjunction with *empirical_formula* and *bigsmiles*. Moreover, the *description* field allows additional descriptions on the component to be provided as free form text. These features are particularly useful for polymer composites (c.f. last example in Section S6 of the Supporting Information). However, whenever possible, users are encouraged to provide chemical structures. In addition, if the concentration or the absolute quantity for a component is known, it can be specified, along with the unit, within the *quantity* and *units* entries. To avoid different synonyms, the units of measure should be specified using the universal resource identifier (URI) of the corresponding object in the Units Ontology.[45] For a mixture component, instead of specifying a chemical structure, the *contents* entry is provided to recursively define its sub-components. Overall, the mixture format is useful when the amount of a component needs to be quantified, including scenarios such as the specification of purity, or the quantification of the concentration of a polymer species within a blend or a solution.

An example illustration of the sub-schema can be found in Figure 4. Within the example, several precursors, including the

a)


```
{  "preamble" : {
      "polymer" : "CC(c1ccccc1){[$][$]CC(C#N)[$],[$]CC(c1ccccc1)[$][$]}Br",
      "pdVersion" : 1.0, "mxfileVersion" : 0.01, "docID" : "doc-xxx.xxx.xxx",  "logs" : [ ... ], "srcs" : [ ... ],
      "network" : [ "[1].[2].[3].[4].[5].[6]>[a]>[7]", "[7]>[b]>[0].[8]" ] },
   "species" : [
      {  "ID" : "[1]", "contents" : [ {  "ID": "[1:1]", "bigsmiles" : "C=CC#N",
                                         "quantity" : "0.60", "units" : "UO_0000013" } ] },
      {  "ID" : "[2]", "contents" : [ {  "ID": "[2:1]", "bigsmiles" : "C=Cc1ccccc1",
                                         "quantity" : "1.0", "units" : "UO_0000013" } ]     },
      {  "ID" : "[3]", "contents" : [ {  "ID": "[3:1]", "bigsmiles" : "CC(c1ccccc1)Br",
                                         "quantity" : "0.040", "units" : "UO_0000013" } ]   },
      {  "ID" : "[4]", "contents" : [ {  "ID": "[4:1]", "bigsmiles" : "[Cu]Br",
                                         "quantity" : "0.060", "units" : "UO_0000013" } ]   },
      {  "ID" : "[5]", "contents" : [ {  "ID": "[5:1]", "bigsmiles" : "c1ccc(nc1)c2ccccn2",
                                         "quantity" : "0.12", "units" : "UO_0000013" } ] },
      {  "ID" : "[6]", "contents" : [ {  "ID": "[6:1]", "quantity" : "1.08", "units" : "UO_0000009",
                                         "contents" : [ {  "ID": "[6:1:1]", "quantity": 99, "units": "UO_0000163",
                                                           "bigsmiles": "O(c1ccccc1)c2ccccc2" } ] }   ]   },
      {  "ID" : "[7]", "contents" : [
         { "ID": "[7:1]", "bigsmiles" : "C=CC#N" }, { "ID": "[7:2]", "bigsmiles" : "C=Cc1ccccc1" },
         { "ID": "[7:3]", "bigsmiles" : "CC(c1ccccc1)Br" }, { "ID": "[7:5]", "bigsmiles" : "c1ccc(nc1)c2ccccn2" },
         { "ID": "[7:6]", "bigsmiles" : "O(c1ccccc1)c2ccccc2" }, { "ID": "[7:4]", "contents" : [
            { "ID" : "[7:4:Cu(I)]", "bigsmiles" : "[Cu]Br" }, { "ID" : "[7:4:Cu(II)]", "bigsmiles" : "Br[Cu]Br" } ] },
         { "ID": "[7:polymer]", "bigsmiles" : "CC(c1ccccc1){[$][$]CC(C#N)[$],[$]CC(c1ccccc1)[$][$]}Br"}
      ] },
      {  "ID" : "[8]", "contents" : [
         { "ID": "[8:1]", "bigsmiles" : "C=CC#N" }, { "ID": "[8:2]", "bigsmiles" : "C=Cc1ccccc1" },
         { "ID": "[8:3]", "bigsmiles" : "CC(c1ccccc1)Br" }, { "ID": "[8:5]", "bigsmiles" : "c1ccc(nc1)c2ccccn2" },
         { "ID": "[8:6]", "bigsmiles" : "O(c1ccccc1)c2ccccc2" }, { "ID": "[8:4]", "contents" : [
            { "ID" : "[8:4:Cu(I)]", "bigsmiles" : "[Cu]Br" }, { "ID" : "[8:4:Cu(II)]", "bigsmiles" : "Br[Cu]Br" } ] },
         { "ID": "[8:polymer]", "bigsmiles" : "CC(c1ccccc1){[$][$]CC(C#N)[$],[$]CC(c1ccccc1)[$][$]}Br"}
      ] },
      {  "ID" : "[0]", "contents" : [
         {  "ID": "[0:polymer]", "bigsmiles" : "CC(c1ccccc1){[$][$]CC(C#N)[$],[$]CC(c1ccccc1)[$][$]}Br",
            "characterization" : { ... }    }
      ] }
   ],
   "transformation" : [ ... ]
}
```


b)


**Figure 4.** Illustrative PolyDAT example (a) on an acrylonitrile styrene copolymer synthesized by atom transfer radical polymerization (ATRP) and subsequently purified and the corresponding reaction scheme (b). Details on polymer characterization and specifications on the transformations (reactions and processing) are provided in later parts of the section.

monomers (acrylonitrile as species [1] and styrene as [2]), initiator (1-phenylethyl bromide [3]), catalysts (copper bromide, [4], and 2,2′-bipyridyl, [5]), and the solvent (diphenyl ether of 99 weight percent purity) are first mixed and polymerized via atom transfer radical polymerization (ATRP) (transformation [a]) and then subsequently purified (transformation [b]) to yield the final acrylonitrile styrene copolymer (species [0]) and the residue solution (species [8]). Note that the residue solution is included in this example for completeness; in practice, the user can retain only those species that are relevant to the purpose of the document. Discussions on how the conversion of a reaction can be encoded will be detailed later in Section 2.4. In this example, the monomers, catalysts, and initiator (species [1] through [5]) are specified as pure substances, with the amount of each reagent given in unit of moles (UO_0000013). This is evident from the *contents* array constituting of only a single component. In contrast, the solvent, diphenyl ether (species [6]), is expressed as a mixture. Within its component object, instead of directly specifying the (Big)SMILES, another *contents* field is included for further specification of the constituent sub-components. Here, only one component, along with its purity in units of weight/weight%, is explicitly specified. Note that the

```
"species" : [
    { "ID" : "[f]",    "contents" : [
        { "ID": "[f:1]", "bigsmiles" : "C=CC#N", "quantity" : "1", "units" : "UO_0010006" },
        { "ID": "[f:2]", "bigsmiles" : "C=Cc1ccccc1", "quantity" : "1.67", "units" : "UO_0010006" },
        { "ID": "[f:3]", "bigsmiles" : "CC(c1ccccc1)Br", "quantity" : "0.067", "units" : "UO_0010006" },
        { "ID": "[f:4]", "bigsmiles" : "[Cu]Br", "quantity" : "0.10", "units" : "UO_0010006" },
        { "ID": "[f:5]", "bigsmiles" : "c1ccc(nc1)c2ccccn2", "quantity" : "0.20", "units" : "UO_0010006" },
        { "ID": "[f:6]", "bigsmiles": "O(c1ccccc1)c2ccccc2" }
    ]  },
    …
]
```


**Figure 5.** A PolyDAT snippet in which species [1] through [6] in the example provided in Figure 4 are encoded as a mixture of six components instead of six distinct species.

unit wt/wt% is denoted by its Units Ontology URI UO_0000163. A list of the corresponding URIs for commonly used units is given in Section S3 in the Supporting Information. The polymer solution (species [7]) obtained upon polymerization and the processed polymer species are encoded with similar syntaxes for mixtures, with the concentrations for some of the components explicitly provided. Note that, in this example, the designation of the pure and mixture reagents is made arbitrarily to illustrate the usage of the schema. In practical scenarios, all the reagents could be considered mixtures due to the impurities within commercial grade reagents. Moreover, while encoding each reagent as distinct species is useful to quantify the extensive amount of each reagent, in practice, polymer chemists may instead prefer to provide intensive quantification such as the mole ratios of the different reagents. In this case, the reagents can be aggregated into a single species, with the relative amount of each reagent indicated by the *quantity* field in association with "UO_0010006" (ratio) as the unit. This usage is illustrated in Figure 5, in which the six reagents of the original example (species [1] through [6]) are aggregated into a single species (species [f]) with six components.

Accompanying the fields that define the chemical identities of the components, the *characterization* field provides additional quantification that further reveals the structural properties of the provided components. PolyDAT's *characterization* section is reserved exclusively for chemical characterization. Here, the term "chemical characterization" is defined in a narrow sense and entails only those characterizations of a polymer that provide direct quantification on the atomic connectivity patterns or the repetitive patterns for recurring sequences within a polymer species. In this context, each characterization datum entry provides information on the probability of observing specific molecular species among the ensemble of molecules that compose the polymer sample; to this end, entries supported within this object are restricted to properties that are directly measurable and extractable from experimental quantifications on the molecular ensemble of a single species or component. Examples of such properties include the molecular weight distribution of polymers as well as other structural features such as tacticity or the composition of polymer chains. Specifically, molecular parameters that are dependent on kinetic models or other chemical models, which involve multiple species and provide relations between multiple molecular ensembles, are not included within this section. Instead, these data are encapsulated under the corresponding transformations. This field is most useful for polymer components for which the BigSMILES structural descriptor does not provide sufficient information to fully resolve the underlying molecular ensemble. In principle, this field is a container object that holds the collection of different characterization data pertinent to the polymer component.

The design of the component object is largely adapted from the Mixfile format developed by Clark et al.[31] Mixfile is a schema that provides a means for encoding the constituents of a mixture. Within a Mixfile entity, the contents of a chemical mixture are specified by a series of *component*s that make up the overall mixture, with each component further described by a *component-obj*. Here, PolyDAT's syntax for *component-obj* closely follows that of the Mixfile format. The core fields of Mixfile, including *quantity* and *units*, as well as their definitions[31] are preserved and reused within PolyDAT. Moreover, while other fields defined within Mixfile, such as *name*, *description*, *synonyms*, *relation*, *ratio*, and other optional reference fields are not explicitly included within the subschema presented, they can also be incorporated if necessary. However, the inclusion of these entries within PolyDAT is optional, and they are omitted in the illustration for brevity. Overall, the major revision to Mixfile is how chemical structures are represented. In PolyDAT, instead of using Molfile or SMILES representations, which are inadequate for polymers, chemical connectivity is delineated by BigSMILES strings as well as additional characterization data to provide support for delineating the structures of polymers.

*2.3.1. Characterization Data.* The *characterization* object provides a container for the collection of different characterization data for a polymer component. Within *characterization* objects, characterization data points are recorded in a structure that closely aligns with how experimentalists characterize polymers, and the schema is designed to encourage the logging of raw or minimally treated data. Notably, while the underlying physicochemical principles of distinct characterization techniques may differ significantly, the nature of the values reported in general fall into one of three categories. Within the *characterization* section, three generic templates, *ratio-obj*, *scalar-obj*, and *vector-obj*, are provided for the logging of distinct characterization data points within these categories.

The first type of characterization, using the *ratio-obj* template, provides measurement of the relative ratios between a set of two or more substructures within the target ensemble. A diverse set of characterization methods fall under this category, including common schemes such as the characterization of tacticity for chiral polymers, the determination of head-to-tail configuration for vinyl polymers, or the accounting of composition of copolymers. In addition, advanced analytical techniques such as the characterization of loop defects within a network using network disassembly spectroscopy (NDS)[47−49] and multi-quantum NMR characterization[50] are also sup-

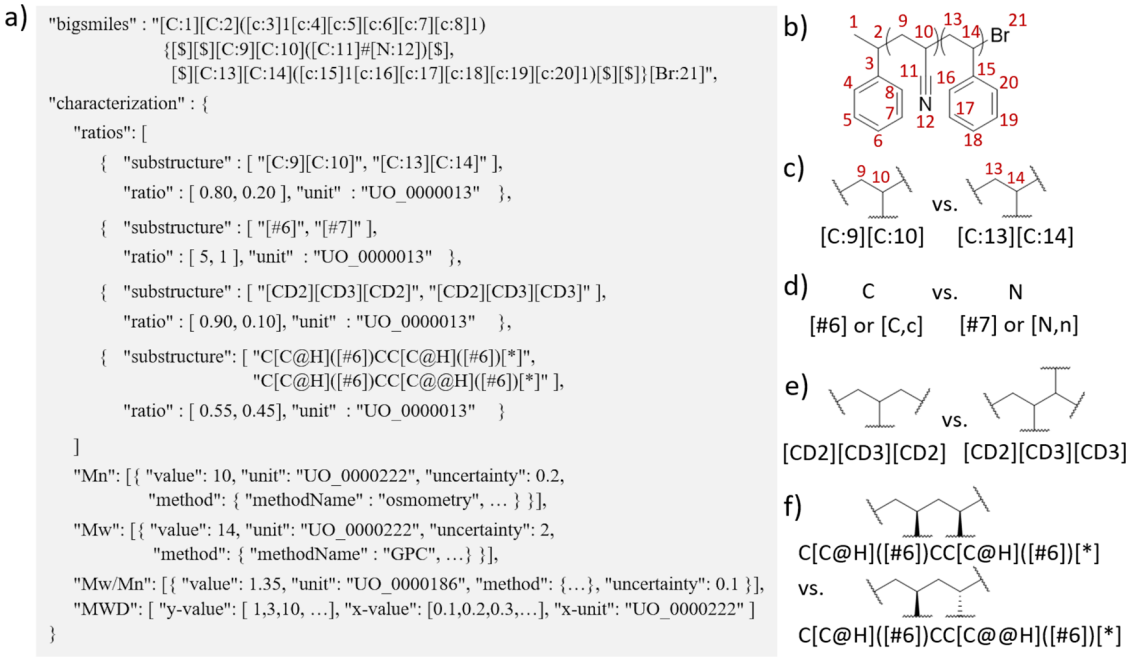



**Figure 6.** (a) Example snippet of a PolyDAT file providing characterization of a poly(styrene-*co*-acrylonitrile), corresponding to the purified species (with ID "[0]") in the example illustrated in Figure 4. (b) Illustration of the polymer along with the (alpha)numeric labels. (c–f) Relevant substructures and corresponding SMARTS strings for specifying the (c) repeat unit composition, (d) elemental composition, (e) head-tail configuration, and (f) tacticity.

ported. Characterization methods associated with this category are jointly collected under the *ratios* section within the *characterization* object. To demonstrate the usage, consider the copolymer, poly(styrene-*co*-acrylonitrile), provided in Figure 4. To quantify its structural features, a list of relevant molecular fragments are first encoded as a *substructure* array of SMILES arbitrary target specification (SMARTS)[51] strings. SMARTS provides a method for specifying substructures of molecules, and a review of different grammatical constructs allowing the user to define a wide variety of relevant substructures can be found in the manual provided by Daylight Chemical Information Systems.[52,53] Then, the relative amount of each molecular fragment is specified in the *ratio* array. Since the ratio is provided in a relative sense, the numeric values need not be normalized. For instance, to specify the composition of the copolymer, the ratio between the two types of carbon backbones, "[C:9][C:10]" and "[C:13][C:14]", can be specified. This example is illustrated in the first item under the *ratios* section in the example provided in Figure 6a,c. Note that, in specifying the molecular fragments, labeled atoms within the SMARTS strings are matched only to the atoms within the species of interest with identical labels, whereas unlabeled atoms can match with any atoms of the same type. As such, the elemental composition of the polymer can be specified by assigning the ratio between the unlabeled atoms. As illustrated in Figure 6d, elemental analysis data can be specified by either using a pattern that explicitly includes both the aliphatic and aromatic atom symbols ("[C,c]" and "[N,n]") or more compactly by specifying the atomic number of the atom ("[#6]" and "[#7]"). By invoking similar strategies, the ratio between successive repeat units in head-to-head and head-to-tail configurations can also be quantified by the relative numbers of tertiary carbons in different molecular environments. As illustrated in Figure 6e, the two relevant states are denoted by the SMARTS strings "[CD2][CD3][CD2]" and "[CD2][CD3][CD3]", where the modifier "D<n>" denotes the number <n> of non-hydrogen neighbors. A more comprehensive manual for other useful SMARTS patterns can be found within the documentation provided by Daylight Chemical Information Systems.[51] Finally, the tacticity of the polymer can also be encoded in a similar fashion using the chirality modifier "@". As sketched in Figure 6f, *meso* diads are represented by the SMARTS string "C[C@H]([#6])CC[C@H]([#6])[*]" and the *racemo* diads "C[C@H]([#6])CC[C@@H]([#6])[*]". Note that "[*]" is a wildcard symbol that matches to any atom. The wildcard atom is used here to explicitly allow the atom trailing the diads to match to either carbon or bromine atoms, whereas the "[#6]" carbon atom is meant to match both the aromatic carbon on the styrene pendant group as well as the aliphatic carbon on the acrylonitrile units.

Similar strategies can be extended to encode other molecular patterns that are far more complicated. In particular, the same scheme could also support many other characterizations that are specific to polymers with non-linear topology. For instance, for a graft polymer (Figure S5a), the grafting efficiency could be specified by providing the ratio of the number of unmodified moieties to the functionalized groups that had went through grafting reaction. Similarly, the branching of polyethylene (Figure S5b) could be quantified by the ratio of secondary carbon to tertiary carbon found within the polymer. Even advanced analytical techniques, such as the NDS,[47−49] which characterizes the number of loop defects within a network by specifying the relative number of junctions in distinct connectivity states (Section S6), are also readily compatible with the schema.

In contrast to the ratio category, the other two categories provide properties that can be recorded as standalone vectors or scalars. In general, these properties are related to the quantification of some distribution of the molecular ensemble

or moments of this distribution. Specifically, characterization of the molecular weight distribution falls into the vector category, whereas characterization of the number or weight average molecular weight and the dispersity of a polymer are represented by scalars. Within PolyDAT, individual characterization data for scalar and vector properties are independently logged under the corresponding entries. Currently, the supported scalar properties include dispersity ("D"), number average molecular weight or molar mass ("Mn"), weight average molecular weight or molar mass ("Mw"), Z average molecular weight or molar mass ("Mz"), number average degree of polymerization ("DPn"), weight average degree of polymerization ("DPw"), Z average degree of polymerization ("DPz"), the skewness of the molecular weight distribution ("skewness"), and the kurtosis of the molecular weight distribution ("kurtosis"). Note that if the number average molecular weight is measured by NMR, since the characterization is done by relating the ratio between substructures, data for such characterization should be provided under the *ratios* entry for consistency. While only nine scalar properties are currently supported, more properties can be readily incorporated. The BigSMILES Project GitHub page[54] contains a live list of all supported measurements. In addition to the scalar moments of the molecular weight distribution, vectoral data such as the molecular weight distribution ("MWD") can also be recorded as vector quantities. Inclusion of full distributions provides the ability to calculate arbitrary moments of the distribution and other distributional properties that may not be possible using the more limited average molar masses that are more commonly reported, enabling the extraction of many additional structural features. An example illustrating the usage of the different entries is provided in Figure 6a.

By providing a combination of multiple characterization techniques across different categories, the schema can provide support for more convoluted characterization. For example, to characterize the number of unsaturated bonds within a polyolefin (Figure S5c), techniques such as the oxidative cleavage of alkenes could be deployed. Once the oxidation is carried out, the combination of the molecular weight distribution and the relative atom counts of oxygen and carbon within the oxidized target polymer as well as the molecular weight distribution of the unsaturated base polymer jointly provide a basis for inferring the fraction of unsaturated olefin bonds within the original polymer.

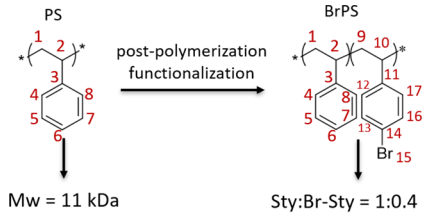
For each property type, more than one data entry can be logged. This feature is included because it is common to have multiple measurements of the same property using multiple tools or instruments. For instance, the moments of molecular weight of a polymer can be measured with gel permeation chromatography (GPC), matrix-assisted laser desorption/ionization (MALDI) mass spectroscopy, and other measurement methods. In general, for each scalar and vector data entry, the measured values, the unit of the value should be provided. Likewise, for ratios, the unit of the ratio, in moles or mass ratio, should be specified. Moreover, for all data entries, the source of the data as well as the method of the measurement should also be specified for data provenance. Notably, for molecular weights obtained via GPC, if the standard-equivalent molecular weight is reported, additional comments on the standard polymer and solvent must be reported along with the characterization method. In addition, uncertainty of the reported values and how the uncertainties are estimated should also be reported if they are available. A complete list detailing the syntax of the supported entries for each data object is given in the Supporting Information. It is essential for the data curator to log all independently measured values and their associated uncertainties as independent entries even if the set of measurements may appear inconsistent as explicitly preserving all data enables more flexibility around the choice of data analytic procedures performed at later stages when the data is used to infer molecular properties and reduce any bias in choice of one measurement over another.

**2.4. The Transformation Section.** The transformation section provides details describing the relationships between different chemical species. Here, a transformation is defined as any physicochemical process that takes as feed one or more species and produce one or more species as product. Similar to the syntax of the species section, the *transformation* section is composed of an array of transformation objects (*transformation-obj*), where each element within the array corresponds to a transformation declared within the preamble. Examples of transformations include chemical reactions, such as simple batch reactions, reactions in a flow reactor, or even reactions in a complex multi-reactor setup. Transformations also include other physicochemical processes such as separation by filtration or distillation. Recording these transformations is critical to completely describing a polymer for two reasons: first, they provide information on synthetic conditions. Second, this framework provides a convenient method for understanding the relationships between the polymer of interest and precursor or post-modified polymers that may have been samples used in various characterization techniques. Because the nature of these transformations can differ significantly, requiring vastly dissimilar schemas to comprehensively describe, PolyDAT specifies a minimal template with only two generic entries: *ID* and *atomMap*:

```
transformation-obj = {
    "ID" : string,
    "atomMap" : [
        [ array of strings ], [ array of strings ], [ array of strings ], …    ]
}
```

The *ID* entry is the string placeholder for the transformation, identical to that declared within the preamble. The *atomMap* entry provides atom-to-atom mapping between the atoms found within the feed species and atoms within the products. The mapping is encoded as an array of string entries. Each tuple is composed of a set of atoms, while the exact atom is specified by concatenating the ID of the (sub-)component and the ID of the atom. For example, to specify the carbon atom on the pendant group of acrylonitrile repeat units in the example illustrated in Figure 6b, the component ID "[0:polymer]" is concatenated with the atom ID "10", yielding overall reference ID "[0:polymer]10". In many cases, the mapping of the atoms is one-to-one. In this case, the corresponding tuple is composed of exactly two elements, each corresponding to an atom in the feed and an atom in the product, respectively. However, the mapping of the atoms need not be one-to-one. In some cases, mappings that are not one-to-one can occur due to the chemistry of the reaction, and the tuple will consist of more than two elements. For instance, when certain reactions are only carried out partially, with some fraction of the atoms in the original chemical environment and others in the converted environment, a one-to-many mapping will result. As an illustration, consider a poly(styrene) that underwent partial bromination, as shown in Figure 7a. In this case, the first backbone carbon atom on the original poly(styrene) (atom

a)




b)

```
{"preamble" : {
    "polymer" : "{[][$]CC(c1ccccc1)[$],[$]CC(c1ccc(Br)cc1)[$][]}",
    "pdVersion" : 1.0, "mixfileVersion" : 0.01, "docID" : "sty-co-br", "logs" : [ … ], "srcs" : [ … ]
    "network" : [ "[1]>[a]>[0]" ]
},
"species": [
  {  "id" : "[1]",  "contents" : [
       {  "ID": "[1:PS]", "bigsmiles" : "{[][$][C:1][C:2]([c:3]1[c:4][c:5][c:6][c:7][c:8]1)[$][]}",
          "characterization" : {  "Mw" : [ {   "value" : 11, "unit" : "UO_0000222"   }  ]  }
  }  ]  },
  {  "id" : "[0]",  "contents" : [
       {   "ID": "[0:BrPS]", "bigsmiles" : "{[][$][C:1][C:2]([c:3]1[c:4][c:5][c:6][c:7][c:8]1)[$],
                                            [$][C:9][C:10]([c:11]1[c:12][c:13][c:14]([Br:15])[c:16][c:17]1)[]}",
          "characterization" : {  "ratios" : [ { "structure" : [ "[C:1][C:2]", "[C:9][C:10]" ],
                                                 "ratio" : [1,0.4], "unit" : "UO_0000013"  }  ]  }
  }  ]  }
],
"transformation" : [
  {  "id" : "[a]",
     "atomMap" : [  ["[1:PS]1","[0:BrPS]1","[0:BrPS]9"], ["[1:PS]2","[0:BrPS]2","[0:BrPS]10"],
                    ["[1:PS]3","[0:BrPS]3","[0:BrPS]11"], ["[1:PS]6","[0:BrPS]6","[0:BrPS]14"],
                    ["[1:PS]4,8","[0:BrPS]4,8","[0:BrPS]12,17"],
                    ["[1:PS]5,7","[0:BrPS]5,7","[0:BrPS]13,16"]   ] }
] }
```

**Figure 7.** Complete example of PolyDAT to illustrate the usage of the transformation template. (a) Illustration of a styrene-brominated styrene copolymer synthesized by bromination of styrene precursor and (b) the corresponding PolyDAT file logging the relevant characterizations and reactions. The transformation section in part (b) provides qualitative knowledge on the atom correspondence in the reactants and products of the functionalization of poly(styrene) illustrated in part (a) through the *atomMap* entries.

"[1:PS]1") corresponds to both the atom on the styrene repeat units (atom "[1:BrPS]1") as well as the atom of the brominated repeat units (atom "[1:BrPS]9") in the product. Therefore, the tuple involving the mentioned atom consists of three elements, as demonstrated in Figure 7b. Likewise, many-to-one mapping can occur when different parts of the polymer unit are converted into the same structure. An example of this is the hydrogenation of unsaturated polyolefins. Many-to-many mappings may also occur when multiple atoms are chemically equivalent due to symmetry. In these cases, instead of presenting these atoms as individual tuple elements, the degeneracy should be indicated by collapsing the indices of the degenerate atoms into a single element. The collapse is done by writing the labels of the degenerate atoms as a comma-delimited list trailing the component ID. For instance, in the example illustrated in Figure 7, the degeneracy of atoms 4 and 8 on the aromatic pendant group will be indicated by writing them as "[1:PS]4,8". Likewise, atoms 5 and 7, as well as the atoms on the post-functionalized polymer, are labeled in the same manner.

For instance, as atoms 4 and 8 in Figure 7a on the unfunctionalized poly(styrene) are chemically equivalent, they map onto the atoms on the product as a pair. This correlation is implicitly captured by lumping the pair of equivalent atoms into the same atom map tuple. Similar treatment is applied to atoms 5 and 7 on the unfunctionalized polymer as well. Furthermore, while encouraged, it is not necessary to explicitly include every species that is involved within chemical reactions. Therefore, atom conservation can be violated within a transformation, and the mapping may only involve a subset of all the atoms found within the feed or the product. This violation in conservation is evident in the provided example as the feed does not contain any bromine atom that corresponds to the bromine found within the product. The final entry, *ratios*, provides a basic handle for specifying quantitative relationships between the reactants and products of a transformation.

Within the base template, only the two fundamental entries common to all transformations are specified. Beyond these entries, the entries within a transformation object are largely unconstrained. Additional fields, such as the temperature and the design of a reaction vessel, or other parameters specific to individual processes, can be appended as needed. To illustrate examples of how the base template can be applied to specific transformations, the Supporting Information contains schema for an ideal batch reactor and for fractionation/precipitation under the assumption of thermodynamic equilibrium. Implementation of these for the styrene–acrylonitrile copolymerization is shown in Figure 8. The snippet in Figure 8

```
"transformation" : [
    {  "id" : "[a]",
       "description" : "isobaric-isothermal-batch-reaction"
       "atomMap" : [  … ],
       "T" : 300,
       "unit-T" : "UO_0000012",
       "P" : 100000,
       "unit-P" : "UO_0000110",
       "conversion" : [
           { "componentID" : "[1:1]", "conversion": 0.9 },
           { "componentID" : "[2:1]", "conversion": 0.85 }
       ]
    }
]
```

**Figure 8.** Illustrative PolyDAT snippets demonstrating the implementation of the styrene–acrylonitrile copolymerization in an isothermal, isobaric batch reactor depicted in Figure 4 (transformation [a]). In this case, within the transformation section, thermodynamic parameters such as the pressure and temperature within the reactor, as well as the conversion of the reactants, are incorporated alongside the base template and presented in the transformation object. Along with the quantification of each substance within the feed provided within the species section illustrated in Figure 4a, the reaction is fully specified. Note that "[1:1]" and "[2:1]" corresponds to the acrylonitrile and styrene monomers, respectively.

demonstrates how the transformation section can be augmented to include information such as the temperature $T$ and pressure $P$, which, together with the initial concentration of individual species delineated within the species section of the feed found in Figure 4a, provides a complete description for the reaction conditions. Furthermore, the combination of the reaction conditions with the conversion for the monomeric species involved fully determines the behavior of a well-mixed ideal batch reactor. Therefore, the adapted transformation object illustrated in Figure 8 provides a convenient template for the quantitative specification of simple batch reactions. Other processes can also be encoded by augmenting the base

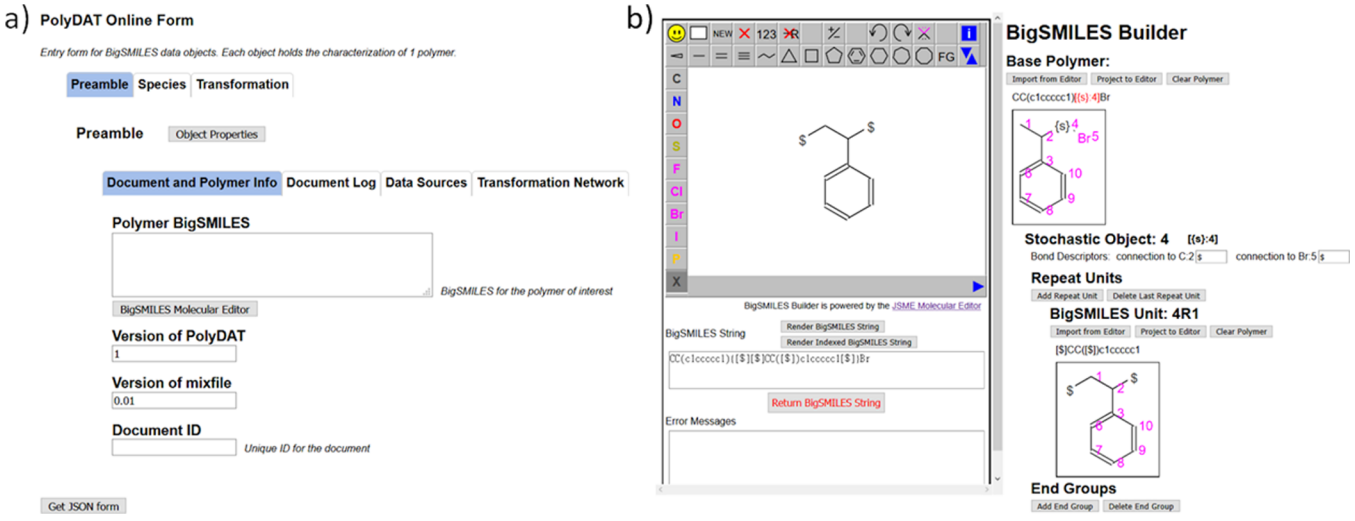



**Figure 9.** Snapshots of the graphical utilities. (a) The PolyDAT web form provides a user-friendly interface for logging data and automatically generates JSON files compliant with the PolyDAT schema. (b) The BigSMILES Builder provides a graphical interface for the specification of polymeric structure and automatically generates the corresponding BigSMILES string.

template in similar ways. For instance, if the critical features such as the thermodynamic parameters are provided, the fractionation process illustrated in Figure 4 can also be captured by within a transformation object. An illustrative example on the precipitation of poly(acrylonitrile) is provided in Figure S9 in the Supporting Information. However, it is important to note that these augmented transformation schemas already contain built-in scientific assumptions regarding the chemistry of a given transformation; the type of information required to specify a transformation will change both as a function of the type of transformation and as a function of the assumptions used. Therefore, PolyDAT retains great flexibility in allowing the user to define the specific model that should be applied to conceptualize each transformation.

Since the transformation section is designed to provide quantitative correlations between the products and the reactants, the section serves as a natural platform for the incorporation of kinetic model and process models and their associated parameters. For instance, apart from reporting reaction conditions, further specifications for a reaction may be provided in the form of a reaction mechanism and the set of associated kinetic parameters, such as the reactivity ratios, the chain transfer rates in a chain copolymerization, as well as the radical concentration predicted from the pseudo-steady state approximation. However, templates for specifying such chemical models and the associated model-dependent parameters are beyond the scope of the base PolyDAT schema, whose major aim is to deliver a model-neutral and assumption-free standard for the quantitative specification of polymer characterization. Therefore, while they offer many merits, templates for model specification will not be provided in this manuscript.

## 3. RESULTS AND DISCUSSION

The schema specified in the previous section provides a standard scheme for systematic categorization and logging of characterization pertinent to a polymer. While only a few simple examples have been illustrated, the proposed schema can be readily applied to much more complicated systems. To illustrate the capacity of the schema, a few examples, including examples on using PolyDAT to record the synthesis of homopolymers, graft polymers, polyelectrolytes, random/block copolymers, and polymer networks, as well as post-polymerization modifications, extracted from the literature have been provided in Section S6 of the Supporting Information.

Although the schema consists of simple components, the complexity of a practical PolyDAT document can quickly become challenging to construct when the network of species and transformations grows through a complex synthetic process. Therefore, to facilitate the adoption of the schema for polymer scientists, a helper program has been implemented and posted on the GitHub page of the project.[54] This utility provides a graphical interface for users to input their characterization data through a web form. A snapshot of the graphical utility is shown in Figure 9a. Once the form is submitted by the user, the input is validated and compared to the schema. If there are required entries that are unfilled, or entries populated with an incorrect data type, then error messages are issued to prompt the user to correct the erroneous entries. Otherwise, if the submitted form passes the validation test, the web form entries are converted into a JSON file compliant with the PolyDAT schema. Under the hood, the web form, the validation program and the JSON converter are generated automatically using the open source JSON Editor[55] and the PolyDAT schema. In addition to the PolyDAT web form, a hierarchical molecular editor, the BigSMILES Builder, which provides a graphical interface for building polymeric structures and converting these polymers into corresponding BigSMILES strings, is also provided. BigSMILES Builder is built on top of the popular JSME molecular editor;[56] a snapshot of the editor is provided in Figure 9b. Since both helper programs are built with JavaScript, they offer cross-platform compatibility. A short tutorial explaining the use of the interface utilities is provided within the Supporting Information.

However, even with the helper program provided, the current ingestion method for PolyDAT is still very basic, and the generation of PolyDAT documents still largely relies on users to provide the inputs in compliance with the proposed syntax. A more refined interactive interface that lessens the burden on the users is currently under development. In particular, ongoing efforts are focused on improving the

BigSMILES Builder, building an interface to allow users to specify the synthetic procedure with graphs, as well as a graphical atom mapping labeling interface. Furthermore, a helper program that can autonomously identify potential atom mapping is also under development. In the long term, the goal is to incorporate natural language processing (NLP) models to extract information from the literature and fully automate the ingestion procedure. Notably, since PolyDAT provides a structured representation of polymeric data, manually curated PolyDAT documents can serve as annotations for training the models that perform information extraction.

The PolyDAT schema is developed to align closely with how experimentalists work with data. Operationally, researchers often exchange sparse datasets that contain information pertinent to only one or a few selected polymers. To reflect this mode of operation, PolyDAT directly aggregates and compiles relevant data for a polymer into a single hierarchical document. This design enables easy dissemination of polymer record in units of individual experiments. Meanwhile, within each document, data are organized according to how characterization is performed experimentally. Therefore, instead of asking for structural properties of the polymer of interest that may be indirectly inferred from measurements performed on other polymers, the schema encourages the user to record characterization data in a format that reflects how the raw data is acquired. For instance, in the styrene-brominated styrene copolymer example illustrated in Figure 7a, even though its molecular weight can be derived from the molecular weight of the poly(styrene) precursor, the molecular weight of the copolymer is not explicitly logged. Instead, the reaction leading to the synthesis of the copolymer is recorded, along with the molecular weight characterization on the precursor, as demonstrated in Figure 7b. There are several advantages to this experiment-centric construct. First, since this multi-species network design closely aligns with how data are extracted from experiments, it provides operational benefits for parties at both ends of the data pipeline. For the creators and curators of the data, minimal additional effort is required to convert their lab notebooks into the proposed format because the schema is simply a digital collection of their experiments. Meanwhile, for the data users, the organization directly reveals the exact set of characterization techniques involved, which improves transparency with regards to how the reported properties of a polymer should be interpreted. In the previous example, if the characterization data had been reduced, and the indirectly inferred molecular weight for the brominated polymer is reported instead, it would have been impossible for a user to tell that a molecular weight characterization had never been performed directly on the brominated polymer. In contrast, by keeping track of the network of all relevant reactions, the context under which individual characterizations are carried out becomes apparent. Furthermore, by associating data with the polymer directly characterized, the recorded data will be assumption-free and minimally biased. This transparency is especially critical for researchers that are less familiar with the experimental techniques from which the physicochemical properties are derived, something that becomes increasingly important for interdisciplinary research between domain experts and data scientists.

While PolyDAT offers a standard schema for the reporting of polymer characterization data, the templates offered within this manuscript are minimal. In particular, the generic templates for the *method* sections within the *ratio-obj*, *scalar-obj*, and *vector-obj* provided in the Supporting Information can be replaced by objects specifically designed for individual characterization techniques. For instance, for molecular weight data extracted from GPC measurements, experimental details such as the type of detectors, the type of column and solvent, the exact model of the equipment used and its calibration, or even the raw GPC data can be incorporated into the *method* section. However, the designing of a template that provides universal compatibility across the realm of existing instruments is far from trivial and well beyond the scope of the current manuscript. Hence, this part of the schema is left mostly unconstrained and flexible.

Within the measurement-centered schema, no data reduction or pre-processing is performed, and each distinct measurement is logged separately. This allows the data creator to provide multiple values of characterization results. For example, multiple chromatographic measurements of molar mass or molar mass characterization by NMR, light scattering, chromatography, and MALDI can all coexist in the same document. Explicitly preserving independent results is especially important in cases where there are differences in values due to uncertainties or biases in the measurements. This feature allows the presentation of the raw data, which provides grounds for further Bayesian analysis that extracts the posterior molecular parameters from experimental data and prior assumptions about the synthetic scheme used for the polymer. Since the original data is provided by the data generator, this enables data users to independently select their preferred priors. For example, in aggregating multiple measurements of the number-average molecular weight performed on different instruments, a researcher who trusts each instrument equally would report the mean of the reported values, whereas a researcher who believes otherwise will bias the result toward the value obtained from the most reliable instrument.

While the proposed PolyDAT characterization schema provides a versatile framework for the encoding of many common characterization data, it also has several limitations. A primary limitation of PolyDAT is that the schema centers around the organization of chemical characterization data that directly inform the molecular connectivity of polymers. Currently, raw data for physical characterization, such as the determination of plateau modulus or glass transition temperature are not directly supported. Furthermore, since the schema encodes the molecular structure of the polymer through a series of direct measurements, its expressive power is limited by the resolution of these characterization experiments. In principle, the structure of any linear copolymer can be captured by incrementally specifying the monadic, diadic, and higher orders relations of the structural units. However, in practice, the resolution of the experimental instruments will introduce a hard limit on the order of structural features that can be accurately resolved. This physical constraint limits the descriptive power of the schema as the document cannot describe a polymer beyond the description provided by the data. In general, the descriptive power of the schema is most severely impaired for structures with long-range correlation, such as gradient linear copolymers or incomplete dendrimers whose junction degrees depend on the number of generations from its core as these structures cannot be approximated well by a truncated series of local structural features.

Furthermore, as the schema is designed to celebrate flexibility and adjacency to experimental practice over conformity, in many cases, the current design may provide

multiple permissible ways of entering the same piece of data. For instance, it has been demonstrated in Figures 4a and 5 that the same precursor mixture can be written in two equally valid manners. This multiplicity makes it more challenging to establish similarity between pairs of PolyDAT documents through juxtaposition. However, as characterization data for polymers are generally sparse, molecular ensembles of polymers are often left underspecified. In many practical cases, this nature renders direct comparison ineffective, and careful interpretation of the provided information through statistical inference is often unavoidable for drawing meaningful comparisons. As such, the current design leans toward providing flexibility and allowing users to specify data in their preferred format to the largest extent. Potential routes to further canonicalize the schema as well as innovative ways of defining and computing the similarity between polymers are being actively pursued.

A potential remedy to the shortcoming of the chemical-characterization-centric schema is the incorporation of predictive models that translate an underlying mechanistic understanding of the chemical system to the knowledge on chemical structures. Such predictive models can involve multiple scales, ranging from process level models, which track the spatiotemporal variation of the concentrations of different species within a reaction vessel, to kinetic models of reaction mechanisms, which specify the relevant reaction networks and the kinetic models associated with individual reactions. As discussed in the methods section, these predictive models can be readily integrated into PolyDAT files under the *transformation* section. To provide a neutral platform, unlike the *characterization* section, the *transformation* section of the schema is largely left unconstrained. While this design ensures that the schema is free of any implicit assumptions that undermines the general applicability of the schema, it also extends significant flexibility that will lead to highly non-standard extensions. Since templates for models must be developed in a case-by-case manner, future revisions of the schema may develop an open framework that attracts community-based efforts in producing templates for different chemical models that are useful and widely accepted.

Finally, while standardized schemas are helpful in collecting polymer data, having a schema does not guarantee the quality of the data. PolyDAT partially address the quality assurance problem by requiring file generators to report the sources of the data. This requirement enables data users to examine the provenance when possible. Incorporation of data sources also allows users to differentiate between high fidelity sources that are actively curated and ordinary sources. In addition to data sources, users are also encouraged to report the uncertainties associated with individual measurements. This record provides additional opportunities to examine the fidelity of the reported data. If multiple measurements of the same property are reported, statistical tools can be applied to check for internal inconsistency and uncover suspicious data points. Furthermore, as many properties are not completely independent, even if only a single measurement is reported for each property, inference could still be carried out. For instance, even if only one dispersity measurement is reported, the value can still be compared to the reported values of the number-average and the weight-average molecular weights. Overall, the explicit inclusion of the uncertainty fields provides a convenient platform for additional data validation procedures as well as room for further statistics-based property estimation with rigorous analytical tools during data assimilation and aggregation.

Overall, PolyDAT provides a standardized format for reporting the characterization and reaction network associated with a polymer species of interest. While it can be used as a standalone document model in a document database that provides encapsulation of chemical characterization, the utility of PolyDAT can be exploited the most when it is used as a descriptive section and embedded within a larger data object model that supports the reporting of additional polymeric data, such as the thermal properties or mechanical properties of polymers. At its core, PolyDAT provides a standard language to specify the chemical "metadata" of polymeric data entries, revealing not only detailed quantification on the chemical structure of the random molecules but also information pertinent to the synthesis/processing of such chemical species, encoded in a graphical format that represents directed acyclic graphs (DAG). These characteristics make PolyDAT ideal to serve as the chemistry-descriptive component in larger data models that involve polymers. To this end, PolyDAT offers a potential resolution to the long-standing challenge of disparate polymer data in polymer informatics and provides the necessary underlying infrastructure to building FAIR (findable, accessible, interoperable, and reusable)[38] digital assets for the polymer community.

## 4. CONCLUSIONS

In this work, the PolyDAT schema for specifying characterization on the molecular structures of polymers is proposed. Unlike most schemas, where only characterization data pertinent to the polymer species of interest are included, PolyDAT utilizes a multi-species format that closely aligns with how experimentalists generate and record data for polymers. This construct allows characterizations to be attached directly to the molecules on which characterization is performed, providing benefits in both data curation and interpretation of the data. In addition, a standardized way of reporting characterization data for a polymer is proposed. Unlike other tabular schemas, which require the curation of a fixed list of relevant properties, PolyDAT categorizes most common types of characterization into three classes, which allows most characterization data to be encoded using one of the generic templates provided. The explicit incorporation of reactions and transformation sections alongside chemical characterizations not only provides a convenient basis for the encoding of crucial correlations between distinct species but also complements the mostly experiment-oriented schema by allowing the incorporation of chemical models. Notably, experimental characterization data and model-dependent parameters are contained in independent sections, explicitly ensuring the base schema to be model-neutral and free of underlying assumptions.

Overall, PolyDAT addresses several critical challenges in polymer informatics. First, by providing a container for the quantitative characterization data, PolyDAT complements existing structural-based polymer representations such as BigSMILES, providing the assignment of a probability or weight to each molecular graph within the ensemble. Next, by offering a standardized schema, PolyDAT provides a universal template for logging characterization and reaction data. By introducing a common interface, such a standard template would significantly lower the barrier to data sharing within the polymer community, motivating the development of a collaborative polymer data repository. In addition, the schema

would also introduce a standard language that serves as a bridge between the polymer community and other informatics and modeling communities, further stimulating the progress of polymer informatics. These advances are likely to be particularly empowering as polymers of increasing molecular complexity, accompanied by access to new function and property spaces, are developed. Ultimately, it is hoped that the proposed schema would serve as a pivotal tool to advance polymer informatics and eventually lead to not only the digitalization of all future polymer data but also the creation of community-wide, large-scale data platforms that resemble initiatives such as the Protein Data Bank or the Cambridge Structural Database.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00028.

Graphical illustration of schema components; detailed specification for the syntax of the Preamble, Species, and Transformation sections; example of units supported in the Units Ontology; examples on the application of PolyDAT to characterizations such as tacticity, regiosequence, and grafting density; sample transformation templates for recording batch reactions and fractionation processes; sample PolyDAT files for nine selected examples; tutorial for PolyDAT Web Form and BigSMILES Builder; and text version copies of PolyDAT snippets presented in figures found in the main text (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Bradley D. Olsen** − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-7272-7140; Email: bdolsen@mit.edu

### Authors

**Tzyy-Shyang Lin** − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

**Nathan J. Rebello** − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-0178-7701

**Haley K. Beech** − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-3276-8578

**Zi Wang** − *Department of Chemistry, Duke University, Durham, North Carolina 27708, United States;* orcid.org/0000-0003-2544-3572

**Bassil El-Zaatari** − *Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States*

**David J. Lundberg** − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

**Jeremiah A. Johnson** − *Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-9157-6491

**Julia A. Kalow** − *Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States;* orcid.org/0000-0002-4449-9566

**Stephen L. Craig** − *Department of Chemistry, Duke University, Durham, North Carolina 27708, United States;* orcid.org/0000-0002-8810-0369

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00028


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS


This work was funded by the Center for the Chemistry of Molecularly Optimized Networks, a National Science Foundation (NSF) Center for Chemical Innovation (CHE-1832256). The authors would like to thank Dr. Debra Audus for her helpful comments.


## ■ REFERENCES


(1) Codd, E. F. Relational Database: A Practical Foundation for Productivity. In *Readings in Artificial Intelligence and Databases*; Elsevier: 1989, pp. 60−68.

(2) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(3) Irwin, J. J.; Shoichet, B. K. ZINC− a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(4) Sussman, J. L.; Lin, D.; Jiang, J.; Manning, N. O.; Prilusky, J.; Ritter, O.; Abola, E. E. Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 1078−1084.

(5) Cochrane, G.; Akhtar, R.; Aldebert, P.; Althorpe, N.; Baldwin, A.; Bates, K.; Bhattacharyya, S.; Bonfield, J.; Bower, L.; Browne, P.; Castro, M. *Nucleic Acids Res.* **2007**, *36*, D5−D12.

(6) Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L. *GenBank. Nucleic Acids Res.* **2007**, *36*, D25−D30.

(7) Sugawara, H.; Ogasawara, O.; Okubo, K.; Gojobori, T.; Tateno, Y. DDBJ with New System and Face. *Nucleic Acids Res.* **2007**, *36*, D22−D24.

(8) Bicerano, J., *Prediction of Polymer Properties, 3rd Edition*. CRC Press: 2002, DOI: 10.1201/9780203910115.

(9) Brandrup, J.; Immergut, E. H.; Grulke, E. A.; Abe, A.; Bloch, D. R., *Polymer Handbook*. Wiley New York: 1999.

(10) *Polymer Property Predictor and Database*. http://pppdb.uchicago.edu/ (accessed Dec 7, 2020).

(11) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 2011; IEEE: *2011*; pp. 22−29.

(12) Mark, J. E., *Physical Properties of Polymers Handbook*. Springer: 2007, DOI: 10.1007/978-0-387-69002-5.

(13) *CHEMnetBASE - Polymers: a Property Database*. http://poly.chemnetbase.com/faces/polymers/PolymerSearch.xhtml (accessed Dec 7, 2020).

(14) *NanoMine*. http://materialsmine.org/nm (accessed Dec 7, 2020).

(15) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **2017**, *6*, 1078−1082.

(16) Jones, R. G.; Wilks, E. S.; Metanomski, W. V.; Kahovec, J.; Hess, M.; Stepto, R.; Kitayama, T., *Compendium of Polymer Terminology and Nomenclature, IUPAC Recommendations 2008*. Royal Society of Chemistry Cambridge: 2009; Vol. 464.

(17) Kahovec, J.; Fox, R. B.; Hatada, K. Nomenclature of Regular Single-Strand Organic Polymers (IUPAC Recommendations 2002). *Pure Appl. Chem.* **2002**, *74*, 1921−1956.
(18) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A.; Jensen, K. F.; Olsen, B. D. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5*, 1523−1531.
(19) *BigSMILES: An open-source effort towards accessible polymer data*. https://olsenlabmit.github.io/BigSMILES/ (accessed Dec 7, 2020).
(20) Cheremisinoff, N. P., *Polymer Characterization: Laboratory Techniques and Analysis*. William Andrew: 1996.
(21) Campbell, D.; Pethrick, R. A.; White, J. R., *Polymer Characterization: Physical Techniques*. CRC press: 2000.
(22) Billingham, N. C., *Molar Mass Measurements in Polymer Science*. Kogan Page London: 1977.
(23) Odian, G., *Principles of Polymerization*. John Wiley & Sons: 2004, DOI: 10.1002/047147875X.
(24) Godwin, A.; Hartenstein, M.; Müller, A. H.; Brocchini, S. Narrow Molecular Weight Distribution Precursors for Polymer−drug Conjugates. *Angew. Chem., Int. Ed.* **2001**, *40*, 594−597.
(25) Neugebauer, D.; Sumerlin, B. S.; Matyjaszewski, K.; Goodhart, B.; Sheiko, S. S. How Dense Are Cylindrical Brushes Grafted from a Multifunctional Macroinitiator? *Polymer* **2004**, *45*, 8173−8179.
(26) Matsunaga, T.; Sakai, T.; Akagi, Y.; Chung, U.-I.; Shibayama, M. Structure Characterization of Tetra-PEG Gel by Small-Angle Neutron Scattering. *Macromolecules* **2009**, *42*, 1344−1351.
(27) Zhao, H.; Wang, Y.; Lin, A.; Hu, B.; Yan, R.; McCusker, J.; Chen, W.; McGuinness, D. L.; Schadler, L.; Brinson, L. C. NanoMine Schema: An Extensible Data Representation for Polymer Nanocomposites. *APL Mater.* **2018**, *6*, 111108.
(28) Huan, T. D.; Mannodi-Kanakkithodi, A.; Kim, C.; Sharma, V.; Pilania, G.; Ramprasad, R. A Polymer Dataset for Accelerated Property Prediction and Design. *Sci. Data* **2016**, *3*, 1−10.
(29) Kaufman, J. G.; Begley, E. F. MatML: A Data Interchange Markup Language. *Adv. Mater. Processes* **2003**, *161*.
(30) Ojala, T. Approaches in Using MatML as a Common Language for Materials Data Exchange. *Data Sci. J.* **2008**, *7*, 179−195.
(31) Clark, A. M.; McEwen, L. R.; Gedeck, P.; Bunin, B. A. Capturing Mixture Composition: An Open Machine-Readable Format for Representing Mixed Substances. *Aust. J. Chem.* **2019**, *11*, 33.
(32) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 928−942.
(33) Murray-Rust, P.; Rzepa, H. S.; Wright, M. Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content. *New J. Chem.* **2001**, *25*, 618−634.
(34) Holliday, G. L.; Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML Vocabulary for Chemical Reactions. *J. Chem. Inf. Model.* **2006**, *46*, 145−157.
(35) Adams, N.; Winter, J.; Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language. *J. Chem. Inf. Model.* **2008**, *48*, 2118−2128.
(36) *GEMD Documentation*. https://citrineinformatics.github.io/gemd-docs/ (accessed Dec 7, 2020).
(37) *FAIR Principles*. https://www.go-fair.org/fair-principles/ (accessed Dec 7, 2020).
(38) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018.
(39) Louzao, I.; Koch, B.; Taresco, V.; Ruiz-Cantu, L.; Irvine, D. J.; Roberts, C. J.; Tuck, C.; Alexander, C.; Hague, R.; Wildman, R.; Alexander, M. R. Identification of Novel "Inks" for 3D Printing Using High-Throughput Screening: Bioresorbable Photocurable Polymers for Controlled Drug Delivery. *ACS Appl. Mater. Interfaces* **2018**, *10*, 6841−6848.
(40) Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H. Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation. *Nat. Rev. Mater.* **2018**, *3*, 5−20.
(41) Barnett, J. W.; Bilchak, C. R.; Wang, Y.; Benicewicz, B. C.; Murdock, L. A.; Bereau, T.; Kumar, S. K. Designing Exceptional Gas-Separation Polymer Membranes Using Machine Learning. *Sci. Adv.* **2020**, *6*, eaaz4301.
(42) *JSON*. https://www.json.org/json-en.html (accessed Dec 7, 2020).
(43) Bray, T.; Paoli, J.; Sperberg-McQueen, C. M.; Maler, E.; Yergeau, F., *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. https://www.w3.org/TR/xml/ (accessed Dec 7, 2020).
(44) Ben-Kiki, O.; Evans, C.; Ingerson, B. *Yaml Ain't Markup Language (YAML™)* version 1.2. https://yaml.org/spec/1.2/spec.html (accessed Dec 7, 2020).
(45) Gkoutos, G. V.; Schofield, P. N.; Hoehndorf, R. The Units Ontology: A Tool for Integrating Units of Measurement in Science. *Database* **2012**, 2012.
(46) *Reaction SMILES and SMIRKS*. https://www.daylight.com/meetings/summerschool01/course/basics/smirks.html (accessed Dec 7, 2020).
(47) Zhou, H.; Woo, J.; Cok, A. M.; Wang, M.; Olsen, B. D.; Johnson, J. A. Counting Primary Loops in Polymer Gels. *Proc. Natl. Acad. Sci.* **2012**, *109*, 19119−19124.
(48) Wang, J.; Lin, T.-S.; Gu, Y.; Wang, R.; Olsen, B. D.; Johnson, J. A. Counting Secondary Loops Is Required for Accurate Prediction of End-Linked Polymer Network Elasticity. *ACS Macro Lett.* **2018**, *7*, 244−249.
(49) Kawamoto, K.; Zhong, M.; Wang, R.; Olsen, B. D.; Johnson, J. A. Loops Versus Branch Functionality in Model Click Hydrogels. *Macromolecules* **2015**, *48*, 8980−8988.
(50) Lange, F.; Schwenke, K.; Kurakazu, M.; Akagi, Y.; Chung, U.-I.; Lang, M.; Sommer, J.-U.; Sakai, T.; Saalwächter, K. Connectivity and Structural Defects in Model Hydrogels: A Combined Proton NMR and Monte Carlo Simulation Study. *Macromolecules* **2011**, *44*, 9666−9674.
(51) *SMARTS - A Language for Describing Molecular Patterns*. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Dec 7, 2020).
(52) *Daylight> SMARTS Tutorial*. https://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html (accessed Dec 7, 2020).
(53) *Daylight> SMARTS Examples*. https://daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html (accessed Dec 7, 2020).
(54) *BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules Without Well-Defined Structures*. https://olsenlabmit.github.io/BigSMILES/ (accessed Dec 7, 2020).
(55) *JSON Editor*. https://github.com/json-editor/json-editor (accessed Dec 7, 2020).
(56) Bienfait, B.; Ertl, P. JSME: A Free Molecule Editor in JavaScript. *Aust. J. Chem.* **2013**, *5*, 24.