PolyDAT: a generic data schema for polymer characterization

Tzyy-Shyang Lin,[†] Nathan J. Rebello,[†] Haley K. Beech,[†] Zi Wang,[§] Bassil El-Zaatari,[‖] David J. Lundberg,[‡] Jeremiah A. Johnson,[‡] Julia A. Kalow,[‖] Stephen L. Craig,[§] Bradley D. Olsen*[,†]

[†]Department of Chemical Engineering and [‡]Department of Chemistry, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States
[§]Department of Chemistry, Duke University, Durham, North Carolina 27708, United States
[‖]Department of Chemistry, Northwestern University, Evanston, Illinois 60208, United States

S1. Overall Schema Structure

The PolyDAT schema is illustrated in Figure S1. As sketched within the illustration, the PolyDAT element consists of a preamble and a data section. In addition, the definition of the log-obj and data-obj objects are also illustrated within Figure S1. Details on each element are provided in later sections of the Supporting Information.

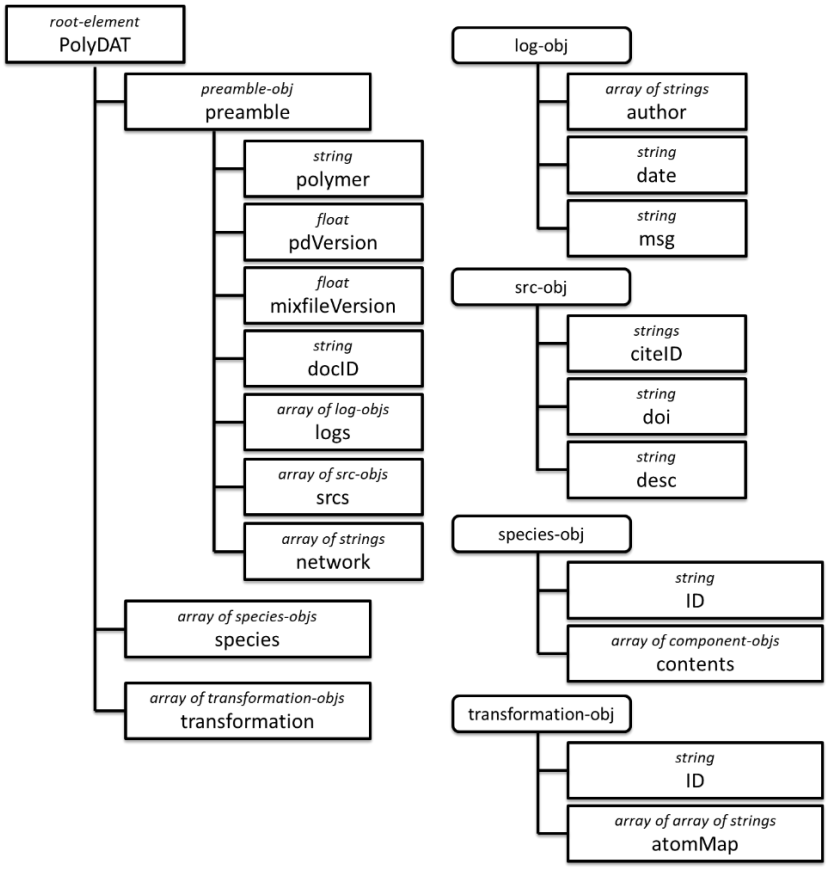


Figure S1. The overall structure of the PolyDAT schema. Each box with a solid border represents a required element. Within each box, the type of the element is denoted in italics, followed by the name of the element. Details on the available data elements are provided in Section S3.

S2. Detailed Specification for the Preamble Section

The preamble section is designed to encapsulate meta-data that provide critical data provenance to the overall PolyDAT object. While additional entries could be incorporated, it is suggested that any preamble includes at least the following entries:

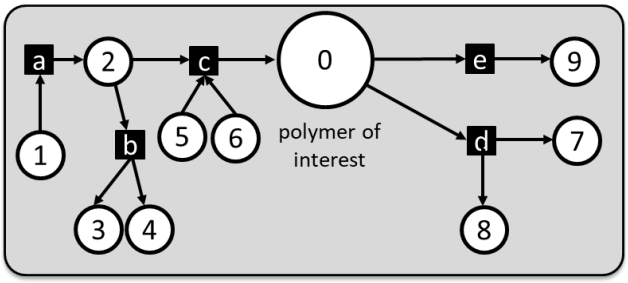
1. "polymer"
    - Type: **string**
    - Required/Optional: **required**
    - Description: A BigSMILES representation that encodes the chemical connectivity of the repeating units, the end groups, and the connectivity patterns between the repeating units and end groups for the polymer of interest. This entry provides a basis for finding the document through structural search. This polymer should be the same as the species with ID [0].
2. "pdVersion"
    - Type: **number**
    - Required/Optional: **required**
    - Description: A number specifying the version of the PolyDAT schema used.
3. "mixfileVersion"
    - Type: **number**
    - Required/Optional: **required**
    - Description: A number specifying the version of the Mixfile format[1] used within the species section of the schema. The Mixfile format provides support for delineating multi-component mixtures.
4. "docID"
    - Type: **string**
    - Required/Optional: **required**
    - Description: A string identifier for the PolyDAT object. This field can serve as the unique document identifier used by organizations within a database for unambiguous referencing of specific documents. While the Digital Object Identifier (DOI) system is recommended, alternative identifier systems may also be deployed.

5. "logs"
    - Type: **array of log-objs**
    - Required/Optional: **required**
    - Description: A revision log, represented by an array of log items. To provide good data provenance, each log item should in the least include the following components:
        i. "author": An array of strings denoting the authors (the authors could be represented with unique identifiers such as the ORCID identifier) of the PolyDAT file.
        ii. "date": String denoting the date in yyyy-mm-dd format
        iii. "msg": Log message that details the changes made to the PolyDAT object.
6. "srcs"
    - Type: **array of src-objs**
    - Required/Optional: **required**
    - Description: An array for the sources of data found within the PolyDAT object. To provide good data provenance, each src item should in the least include the following components:
        i. "citeID": A string label for this specific data source. This label/ID can be used to cite this piece of data within the characterization data section.
        ii. "srcID": A string identifier that provides unambiguous reference to the source of the data. For published data, the Digital Object Identifier System identifier is the preferred choice for indicating the data source. For other data sources that do not have DOI, alternative identifiers can be used instead.
        iii. "desc": A description of the data source.
7. "network"
    - Type: **array of strings**
    - Required/Optional: **required**
    - Description: The field specifies the species and transformations found within the later sections of the PolyDAT file. The number of elements found within the array should equal the number of transformations found in the transformation section. Each element 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:

[feed1].[feed2]. … >[transformation]>[prod1].[prod2]. …

where "[feed1]", "[feed2]", … corresponds to the IDs for the species found within the feed, and "[prod1]", "[prod2]" corresponds to the IDs for the species found within the feed, and "[transformation]" corresponds to the ID of the transformation. Note that the ID for the species and the transformations can be constructed of any alphanumeric strings enclosed within a pair of brackets, with the exception of the polymer species of interest, which must be labelled as "[0]".

A practical example of the preamble, encoding the following reaction network,



is illustrated in the following example

```
{
    "preamble" : {
          "polymer" : "[H]{[>][<]NCC(=O)[>][<]}O",
          "pdVersion" : 1.0,
          "mxfileVersion" : 0.01,
          "docID" : "https://doi.org/10.1000/182",
          "logs" : [
                    {  "author" : [ "ORCID:https://orcid.org/0000-0002-7272-7140" ],
                       "date" : "2020-03-20",
                       "msg" : "document first created" },
                    {  "author" : [ "ORCID:https://orcid.org/0000-0002-7272-7140" ],
                       "date" : "2020-04-20",
                       "msg" : "document revised" }
          ],
          "srcs" : [
                    {  "citeID" : "olsen2019",
                       "doi" : "https://doi.org/10.1021/acscentsci.9b00476",
                       "desc" : "BigSMILES paper" }
          ],
```

… cont'd next page …

```
        "network" : [
                "[1]>[a]>[2]",
                "[2]>[b]>[3].[4]",
                "[2].[5].[6]>[c]>[0]"],
                "[0]>[e]>[9]",
                "[0]>[d]>[7].[8]"
        ]
    },
    "species" : [ ... ],
    "transformation" : [ ... ]
}
```

S3. Detailed Specification for the Species Section

The *species* entry corresponds to an array of different species. Each species is denoted by a species-obj, which is composed of an ID (that matches that provided within the preamble) and a *contents* array composed of *component-obj* elements that further specifies the content of the species:

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

Each *component-obj* element provides specification of a component within the species. The core entries for describing a component are illustrated in the sub-schema shown below:

```
component-obj = {
    "ID" : string,
    "bigsmiles" : string,
    "contents" : array of component-obj,
    "quantity" : number(s),
    "units" : 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
    }
}
```

The *component-obj* is composed of four types of information: (1) the identifier of the specific component, (2) the chemical content of the component, (3) the concentration or the amount of the component, and (4) chemical characterizations associated with the component. The chemical content is either denoted by the BigSMILES of the substance (for a pure component or a polymer species that could be encoded using a single BigSMILES string), or, for a mixture, defined using another *contents* array that provides the list of sub-components. The unit of measurement for the concentration or the amount of the component can be denoted using the universal resource identifier (URI) for the corresponding unit defined within the Units Ontology[2]. URIs for common units are illustrated in the following table

Category	Units	URI
Base Unit	dimensionless unit	UO_0000186
	mass unit	UO_0000002
	ratio	UO_0000190
Substance Unit	mole	UO_0000013
	millimole	UO_0000040
	micromole	UO_0000039
	nanomole	UO_0000041
	picomole	UO_0000042
Mass Unit	Da	UO_0000221
	kDa	UO_0000222
	kg	UO_0000009
	g	UO_0000021
	mg	UO_0000022
	μg	UO_0000023
Volume Unit	L	UO_0000099
	mL	UO_0000098
	cm^3	UO_0000097
	μL	UO_0000101
	nL	UO_0000102
Temperature Unit	K	UO_0000012
	ºC	UO_0000027
Concentration Unit	w/w%	UO_0000163
	w/v%	UO_0000164
	v/v%	UO_0000165
	mol/mol%	UO_0000076
	mol/L	UO_0000062
	mmol/L	UO_0000063
	μmol/L	UO_0000064
	nmol/L	UO_0000065
	pmol/L	UO_0000066

	mol/kg	UO_0000068
	mmol/kg	UO_0000069
	μmol/kg	UO_0000070
	nmol/kg	UO_0000071
	pmol/kg	UO_0000072
	pH	UO_0000196
Molar Mass Unit	g/mol	UO_0000088
	kg/mol	UO_0000087
Density Unit	g/L	UO_0000175
	mg/L	UO_0000273
	μg/L	UO_0000275
	kg/L	UO_0000174
Pressure Unit	Pa	UO_0000110
Time Unit	s	UO_0000010
	h	UO_0000032
	m	UO_0000031
	ms	UO_0000028
	μs	UO_0000029
	ns	UO_0000150
	half life	UO_0000152

A full list the uniform resource identifier (URI) for all supported units can be found at the following website:

https://bioportal.bioontology.org/ontologies/UO/?p=classes&conceptid=root.

The design of the component object is adapted from the Mixfile format developed by Clark et. al.[1] The syntax for *component-obj* closely follows that of the Mixfile format. Apart from the core fields such as *quantity* and *units*, other fields defined within the Mixfile schema can also be used within PolyDAT. Therefore, field including *name*, *description*, *synonyms*, *relation*, *ratio* and other optional reference fields can also be incorporated if necessary. The *name* and *description* fields can be useful if the component is not easily characterized by a chemical structure. Beyond Mixfile fields, additional fields such as the *empirical_formula* can be also be incorporated to denote the composition of material. Here, *empirical_formula* denotes a composition represented as a chemical formula string, in which the order and grouping of the elements is ignored. The inclusion of these entries within PolyDAT is optional. Moreover, additional fields can be incorporated as well if necessary. However, it should be noted that in PolyDAT, Molfile, InChI or SMILES representations are not supported formats for specifying the chemical

connectivity. Only BigSMILES strings are allowed to guarantee the support for polymers. The sub-schema with a full list of available entries are illustrated in the following sub-schema:

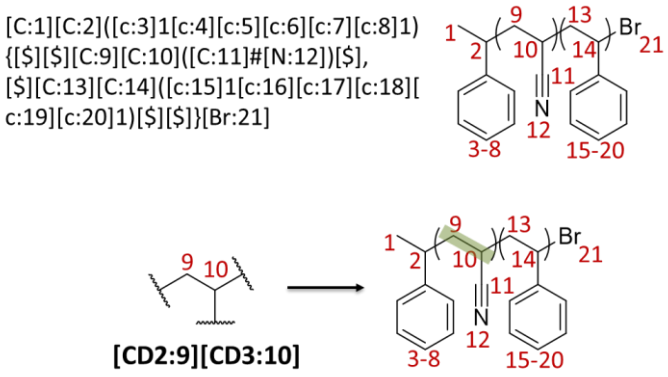
```
component-obj = {
    "ID" : string,
    "bigsmiles" : string,
    "contents" : array of component-obj,
    "empirical_formula" : string,

    "name" : string,
    "description" : string,
    "synonyms" : string,

    "quantity" : number(s),
    "units" : string,
    "relation" : two numbers,
    "ratio" : string,

    "characterization" : { … }
}
```

The *characterization* object provides a container for the collection of different characterization data for a polymer component. Within the *characterization* section, characterization data are categorized into three categories, and three generic templates, *ratio-obj*, *scalar-obj* and *vector-obj*, are provided for the logging of distinct characterization data points within these categories. Several characterizations currently supported within PolyDAT are included within the first *component-obj* sub-schema. Note that the list of the supported characterizations may be updated in future revisions of PolyDAT. For the most up-to-date list of supported characterizations, please refer to the GitHub page for the project (The PolyDAT data schema section within https://olsenlabmit.github.io/BigSMILES/).

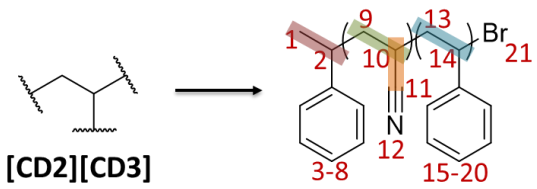
S3.1 Syntax for ratio-obj

The first template, *ratio-obj*, provides a generic format for recording characterization data that essentially measures the relative ratios between a set of two or more substructures within the target ensemble. For the ratios data entries, each characterization data entry should consist of the following entries:

1. "substructure":
    - Type: **array of (string** or **array of strings)**
    - Required/Optional: **required**
    - Description: An array of substructures written in SMARTS syntax that indicate different molecular substructures that could be found within the target polymer. Atoms within the SMARTS string may be labelled. If the label for an atom is provided, the provided SMARTS pattern will only match to atoms that have the identical label. For example, In the example illustrated below, if the labelled pattern "[CD2:9][CD3:10]" is specified, only the vinyl carbons on the acrylonitrile repeat units are matched:

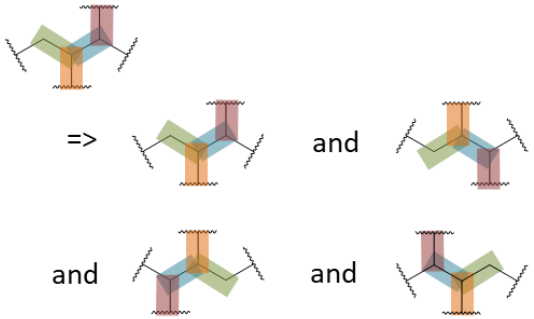


However, if no labels are provided, then all instances of a secondary aliphatic carbon connected to a tertiary aliphatic carbon are considered to be matching, and the pattern "[CD2][CD3]" represents the union of the collection of these substructures:



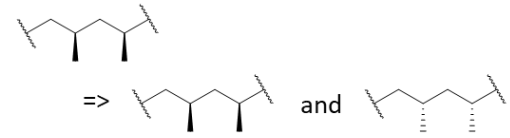
Moreover, in considering the match of a substructure, all substructures isomorphic to the provided pattern upon inversion, reflection, as well as rotation are all implicitly considered and included. Therefore, the example pattern provided within Figure 4e within the main text implicitly includes the four substructures:

[CD2][CD3][CD3]

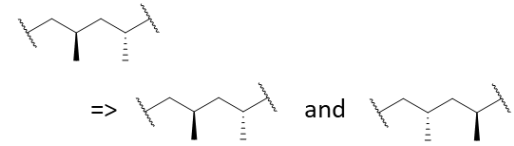


Similarly, the examples in Figure 4f also denote the two possible stereochemical configurations for the raceme and meso diads:

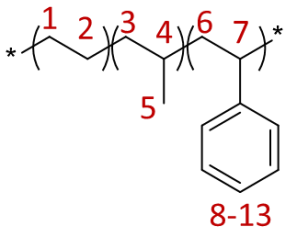
C[C@H]([#6])CC[C@H]([#6])[*]



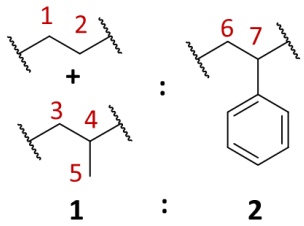
C[C@H]([#6])CC[C@@H]([#6])[*]



Furthermore, within the *substructure* array, multiple SMARTS may be collected into an array element to represent the union of the substructures. In this case, any substructure found within the polymer matches to any pattern within the set. For example, for the following ethylene-propylene-styrene copolymer,



if the total amount of the ethylene and propylene repeat units combined constitutes of one third of the total repeat units



the corresponding entry within the *ratios* section should read


```
{
   "substructure" : [ ["[C:1][C:2]", "[C:3][C:4]([C:5])"], "[C:6][C:7](c1ccccc1)" ],
   "ratio" : [1,2],
   "units" : "mole"
}
```


where the ethylene and propylene repeat units are lumped into an entry within the *substructure* array.

2. "ratio"
    - Type: **array numbers**
    - Required/Optional: **required**
    - Description: An array of quantifying the relative amounts of the substructures specified within the *substructure* array found within the species. The size of the array must be identical to that of the *substructure* array.
3. "unit"
    - Type: **string**
    - Required/Optional: **required**
    - Description: The units for the ratio provided. Available options are:
        1. "UO_0000013": ratio measured in moles
        2. "UO_0000002": ratio measured in mass
4. "src"
    - Type: **string**
    - Required/Optional: **optional**
    - Description: A string indicating the source of the data. The entry should be the *citeID* for one of the entries found in the *srcs* array.

5. "method"
    - Type: **method-obj** or **string**
    - Required/Optional: **optional**
    - Description: An object indicating how the measurement is performed. The entries of the **method-obj** object is meant for specifying the experimental "metadata" associated with the characterization data reported in the *ratio* section. The object should always contain the following entry:
        1. "methodName"
            - Type: **string**
            - Required/Optional: **required**
            - Description: A string indicating the name of the measurement involved. Examples include: GPC, MALDI, NMR, … etc.

      Other than the *methodName* field, additional fields specific to the characterization technique can also be incorporated into the *method-obj* object. In particular, raw data, such as raw NMR data, can also be encapsulated within this section. Owning to the diversity in characterization techniques, the syntax for **method-obj** is left largely unconstrained so that it can be flexibly adapted for different techniques. If **string** is used in place of the **method-obj**, then the string should be the "methodName" string.
6. "uncertainty"
    - Type: **number** or **array of numbers**
    - Required/Optional: **optional**
    - Description: A number or an array of numbers indicating the uncertainty in the reported ratios, in the same unit as that provided in *unit*. If an array is provided, the size of the array must be identical to that of the *substructure* and *ratio* arrays, with each array element corresponding to the uncertainty for each entry provided in *ratio*. If a scalar is specified, the value is interpreted as the uncertainty for every entry within *ratio*.
7. "uncertainty_src"
    - Type: **string**
    - Required/Optional: **optional**
    - Description: A string detailing how the uncertainty is estimated.

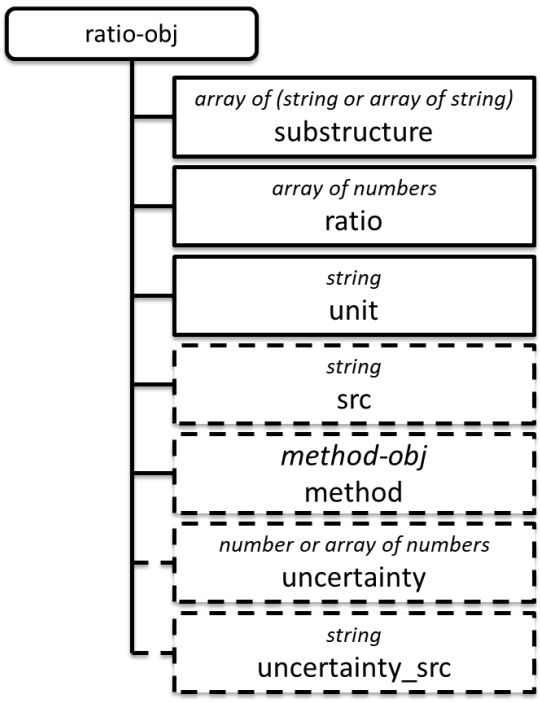


Figure S2. Illustration of the members of the *ratio-obj* object. Solid boxes denoted required fields, whereas dashed boxes are optional.

S3.2 Syntax for scalar-obj

The second template, *scalar-obj*, provides a generic format for recording characterization data that are scalar values. Currently, the supported scalar properties include dispersity ("Mw/Mn"), 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"). Each independent characterization should be logged as separate array element under the entry for the corresponding property. Each *scalar-obj* object is composed of the following fields:

1. "value":
    - Type: **number**
    - Required/Optional: **required**
    - Description: The numeric value of the property of interest
2. "unit"
    - Type: **string**
    - Required/Optional: **required**
    - Description: The units for the value provided. Available options are:

Property	Supported Units	Notes
Mw/Mn	UO_0000186	dim-less unit
Mn	UO_0000221 or UO_0000222 or UO_0000088 or UO_0000087	Da or kDa or g/mol or kg/mol
Mw	UO_0000221 or UO_0000222 or UO_0000088 or UO_0000087	Da or kDa or g/mol or kg/mol
Mz	UO_0000221 or UO_0000222 or UO_0000088 or UO_0000087	Da or kDa or g/mol or kg/mol
DPn	UO_0000186	dim-less unit
DPw	UO_0000186	dim-less unit
DPz	UO_0000186	dim-less unit
skewness	UO_0000186	dim-less unit
kurtosis	UO_0000186	dim-less unit

3. "src"
    - Type: **string**
    - Required/Optional: **optional**
    - Description: A string indicating the source of the data. The entry should be the *citeID* for one of the entries found in the *srcs* array.
4. "method"
    - Type: **method-obj**
    - Required/Optional: **optional**
    - Description: An object indicating how the measurement is performed. The entries of the **method-obj** object is meant for specifying the experimental "metadata" associated with the characterization data reported in the *value* section. The object should always contain the following entry:
        1. "methodName"
            - Type: **string**
            - Required/Optional: **required**
            - Description: A string indicating the name of the measurement involved. Examples include: GPC, MALDI, NMR, … etc.

      Other than the *methodName* field, additional fields specific to the characterization technique can also be incorporated into the *method-obj* object. In particular, raw data, such as raw GPC data, as well as the specification of the instrument used for measurement, can also be encapsulated within this section.
5. "uncertainty"
    - Type: **array of numbers**
    - Required/Optional: **optional**
    - Description: An array of numbers indicating the uncertainty in the reported ratios, in the same unit as that provided in *unit*. The size of the array must be identical to that of the *substructure* and *ratio* arrays.
6. "uncertainty_src"
    - Type: **string**
    - Required/Optional: **optional**
    - Description: A string detailing how the uncertainty is estimated.

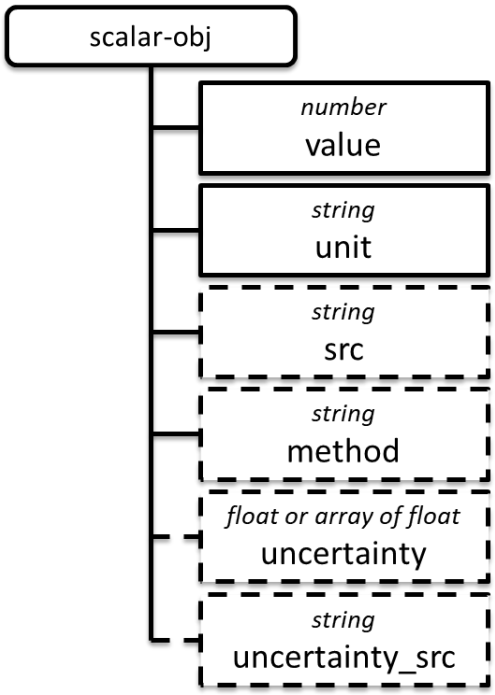


Figure S3. Illustration of the members of the *scalar-obj* object. Solid boxes denoted required fields, whereas dashed boxes are optional.

S3.3 Syntax for vector-obj

The third template, *vector-obj*, provides a generic format for recording characterization data that are vectors. The currently supported vector properties include only the molecular weight distribution ("MWD"). Similar to the *scalar-obj*, the *vector-obj* object is composed of the fields corresponding to measured numeric values, the source, the uncertainty, and the method. However, for distributional data, both the values of the x- and y-axes must be recorded.

1. "y-value":
    - Type: **array of number**
    - Required/Optional: **required**
    - Description: The intensities of the measured data, in arbitrary unit.
2. "x-value":
    - Type: **array of number**
    - Required/Optional: **required**
    - Description: The numeric values of the property corresponding to the entries within *y-value*. For GPC, the property corresponds to the retention time; for MALDI, the mass to charge ratio, and the and molecular weight for MWD.
3. "x-unit"
    - Type: **string**
    - Required/Optional: **required**
    - Description: The unit of measurement associated with the *x-value*. Available options are:
        - For MWD:
            1. "UO_0000221": Da
            2. "UO_0000222": kDa
            3. "UO_0000088": g/mol
            4. "UO_0000087": kg/mol
4. "src"
    - Type: **string**
    - Required/Optional: **optional**
    - Description: A string indicating the source of the data. The entry should be the *citeID* for one of the entries found in the *srcs* array.

5. "method"
    - Type: **method-obj**
    - Required/Optional: **optional**
    - Description: An object indicating how the measurement is performed. The entries of the **method-obj** object is meant for specifying the experimental "metadata" associated with the characterization data reported in the *y-value-x-value* sections. The object should always contain the following entry:
        1. "methodName"
            - Type: **string**
            - Required/Optional: **required**
            - Description: A string indicating the name of the measurement involved. Examples include: GPC, MALDI, … etc.

        Other than the *methodName* field, additional fields specific to the characterization technique can also be incorporated into the *method-obj* object. In particular, raw data, such as raw GPC data, as well as the specification of the instrument used for measurement, can also be encapsulated within this section.
6. "uncertainty"
    - Type: **number** or **array of numbers**
    - Required/Optional: **optional**
    - Description: A number or an array of numbers indicating the uncertainty in the reported y-values, in the same unit. If an array is provided, the size of the array must be identical to that of the *value* array, with each array element corresponding to the uncertainty for each entry provided in *value*. If a scalar is specified, the value is interpreted as the uncertainty for every entry within *value*.
7. "uncertainty_src"
    - Type: **string**
    - Required/Optional: **optional**
    - Description: A string detailing how the uncertainty is estimated.

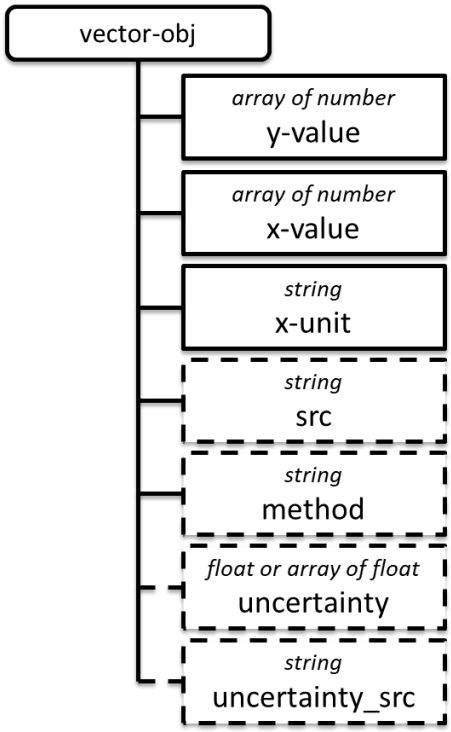


Figure S4. Illustration of the members of the *scalar-obj* object. Solid boxes denoted required fields, whereas dashed boxes are optional.

S4. Examples on the Application of PolyDAT on Selected Characterizations

Beyond the basic examples illustrated within the main text, several additional examples are provided here to further illustrate the application of the schema in different cases. Beyond the characterization of linear structure illustrated in the main text, the *ratio* section can be readily applied to quantify structural features of nonlinear polymers. For instance, consider the graft polymer sketched in Figure S5a. The grafting efficiency of the polymer can be quantified by specifying the relative number of poly(glycidyl azide) repeat units that have yet to undergone the click reaction and those had already been grafted. A full illustrative example on using PolyDAT to report the characterization of a poly(glycidyl azide-*graft*-polyethylene) synthesized by grafting to is shown in Figure S6a, with the corresponding reaction schematically illustrated in Figure S6b.

Like polymer grafts, the branching of a branched poly(ethylene) can also be quantified in a similar fashion. Consider the example illustrated in Figure S5b, where the branched poly(ethylene) consists of a linear repeat unit with two secondary carbons and a branched unit with a tertiary carbon and a secondary carbon. Therefore, the frequency of encountering branching sites can be quantified by specifying the ratio between the carbon atoms with different connectivity states (secondary vs. tertiary). In addition to direct quantification of substructures within a polymer, multiple characterizations across different polymers can also be combined to provide further quantification. Furthermore, characterization across different categories can also be combined to provide support for more convoluted characterization. For instance, oxidative cleavage of alkenes is often deployed to characterize the amount of unsaturated bonds within a polyolefin. As illustrated in Figure S5c, the combination of the molecular weight distribution and the relative atom counts of oxygen and carbon within the oxidized polymer as well as the molecular weight distribution of the unsaturated polymer jointly provide a basis for inferring the fraction of unsaturated olefin bonds within the original polymer.

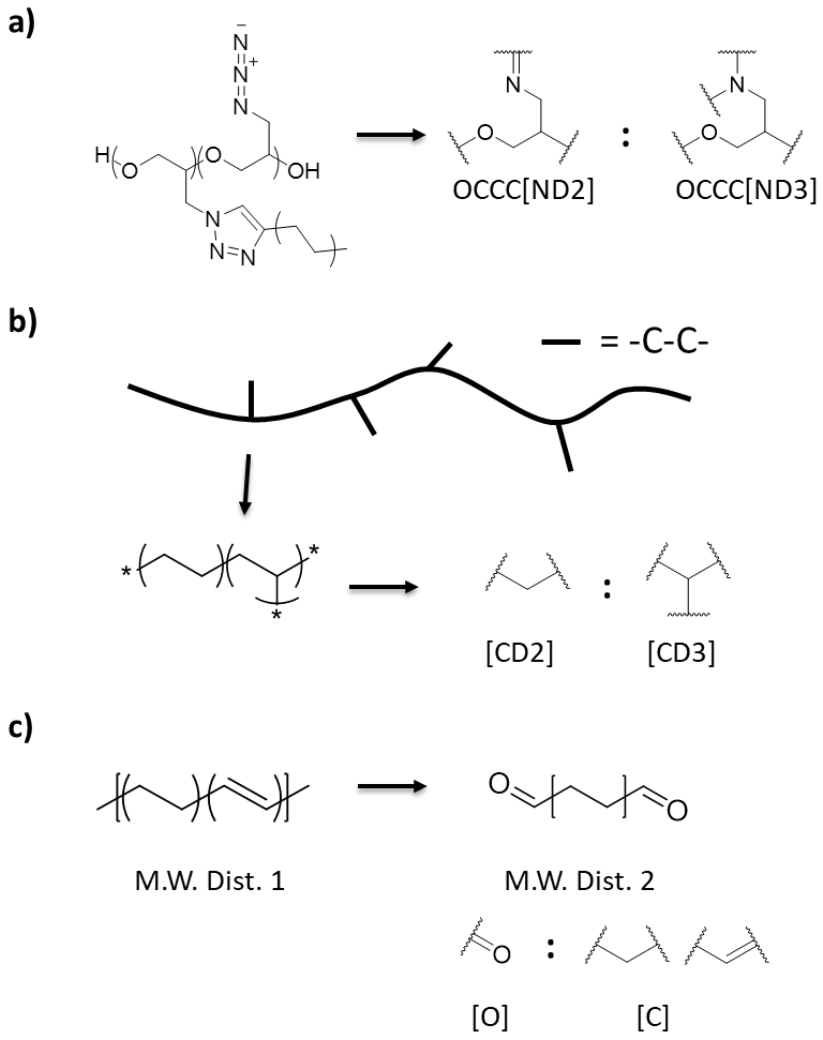


Figure S5. Illustration of using the *ratio* section to quantify the structure of polymers. (a) Quantifying the grafting efficiency of a poly(glycidyl azide) polymer grafted with azide grafts. The distinction between the grafted and ungrafted repeat units can be distinguished by the connectivity on the nitrogen atom. For ungrafted units, the first nitrogen atom encountered from the ethylene glycol repeat has two neighboring heavy atoms, whereas the corresponding nitrogen atom on grafted units are tertiary. (b) The branching of a poly(ethylene) can be quantified by specifying the ratio between the number of secondary carbons and tertiary carbons. (c) Combining multiple characterizations across difference species may provide further knowledge on the structure for the polymer of interest.

a)


```
{
   "preamble" : {
      "polymer" : "{[>][<]OCC(CN=[N+]=[N-])[>],
                     [<]OCC(CN1N=NC(=C1){[$][$]CC[$][$]}C)[>][<]}O",
      "pdVersion" : 1.0,
      "mxfileVersion" : 0.01,
      "docId" : "doc-PolyDAT-S5-a",
      "logs" : [   {  "author" : [ "ORCID:https://orcid.org/0000-0002-7272-7140" ],
                     "date" : "2020-05-04",
                     "msg" : "example first created" }        ],
      "srcs" : [ … ],
      "network" : [ "[azide].[alkyne]>[a]>[0]"]
   },
   "species": [
      {
         "id" : "[azide]",
         "contents" : [
            {   "ID": "[azide:PGA]" ,
               "bigsmiles" : "{[>][<][O:1][C:2][C:3]([C:4][N:5]=[N+:6]=[N-:7])[>][<]}[O:8]",
               "name" : "poly(glycidyl azide)",
               "characterization" : {
                  "Mn" : [  {   "value": 2.5, "unit": "UO_0000222",
                                 "method": { "methodName" : "osmometry"}, "uncertainty": 0.5 },
                              {    "value": 2.4, "unit": "UO_0000222",
                                 "method": { "methodName" : "GPC"}, "uncertainty": 0.4 }          ],
                  "MWD" : [{ "y-value": [4.1, 9.8, 18.2, 22.4, 20.0, 14.9, 12.2, 9.8],
                                 "x-value": [2400,2440,2480,2520,2560,2600,2640,2680],
                                 "x-unit" : "UO_0000222",
                                 "method": { "methodName": "GPC", "standards": "polystyrene" }
                                 }   ]
               }   }   ]
      },
```


… cont'd next page …

```
        {   "id" : "[alkyne]",
            "contents" : [
                {   "ID": "[alkyne:PE]" ,
                    "name" : "alkyne terminated poly(ethylene)",
                    "bigsmiles" : "[C:1]#[C:2]{[$][$][C:3][C:4][$][$]}[C:5]",
                    "characterization" : {
                        "Mn" : [  {   "value": 1.5, "unit": "UO_0000222", "uncertainty": 0.4 }  ],
                        "D" : [     {   "value": 1.1, "unit": "UO_0000186"        }          ]
        }   }   ]   },
        {   "id" : "[0]",
            "contents" : [
                {   "ID": "[0:PGA-g-PE]" ,
                    "name" : "poly(glycidyl azide-graft-polyethylene)",
                    "bigsmiles" : "{[>][<][O:1][C:2][C:3]([C:4][N:5]=[N+:6]=[N-:7])[>][<]}
                                    [<][O:11][C:12][C:13]([C:14][N:15]1[N:16]=[N:17][C:18](=[C:19]1)
                                    {[$][$][C:20][C:21][$][$]}[C:22])[>][<]}[O:8]",
                    "characterization" : {
                        "ratios" : [    {   "substructure": [ "[O:1][C:2][C:3][C:4][ND2:5]",
                                                              "[O:11][C:12][C:13][C:14][ND3:15]" ],
                                        "ratio" :[ 0.1, 0.9 ], "unit" : "UO_0000013"     }                    ]
        }   }   ]   },
    ],
    "transformation" : [     {
            "ID": "[a]", "description" : "azide-alkyne graft-to",
            "atomMap" : [
                [ "[azide:PGA]1", "[0:PGA-g-PE]1", "[0:PGA-g-PE]11" ],
                [ "[azide:PGA]2", "[0:PGA-g-PE]2", "[0:PGA-g-PE]12" ],
                [ "[azide:PGA]3", "[0:PGA-g-PE]3", "[0:PGA-g-PE]13" ],
                [ "[azide:PGA]4", "[0:PGA-g-PE]4", "[0:PGA-g-PE]14" ],
                [ "[azide:PGA]5", "[0:PGA-g-PE]5", "[0:PGA-g-PE]15" ],
                [ "[azide:PGA]6", "[0:PGA-g-PE]6", "[0:PGA-g-PE]16" ],
                [ "[azide:PGA]7", "[0:PGA-g-PE]7", "[0:PGA-g-PE]17" ],
                [ "[azide:PGA]8", "[0:PGA-g-PE]8" ], [ "[alkyne:PE]1", "[0:PGA-g-PE]19" ],
                [ "[alkyne:PE]2", "[0:PGA-g-PE]18" ], [ "[alkyne:PE]3", "[0:PGA-g-PE]20" ],
                [ "[alkyne:PE]4", "[0:PGA-g-PE]21" ], [ "[alkyne:PE]5", "[0:PGA-g-PE]22" ]
            ]
        }
    }
}
```

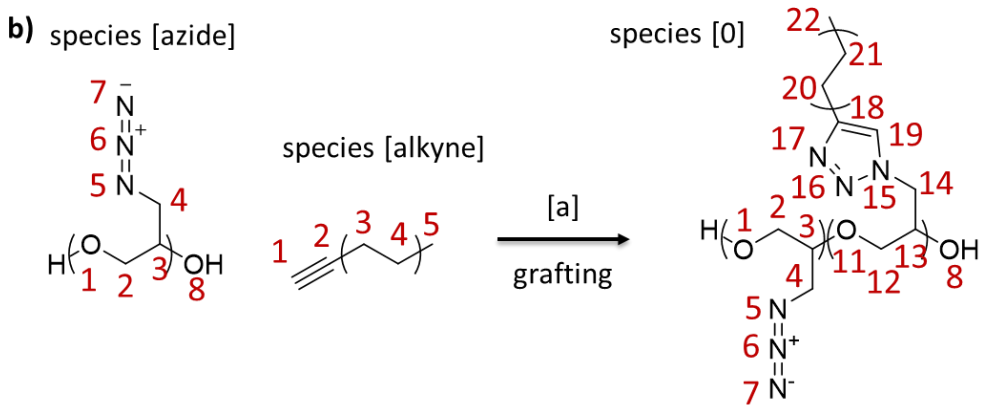


Figure S6. (a) Example illustrating the use of the PolyDAT to record characterization of a graft polymer. As illustrated in part (b), the graft polymer is synthesized by grafting an alkyne terminated polyethylene to poly(glycidyl azide).

S5. Detailed Specification for the Transformation Section

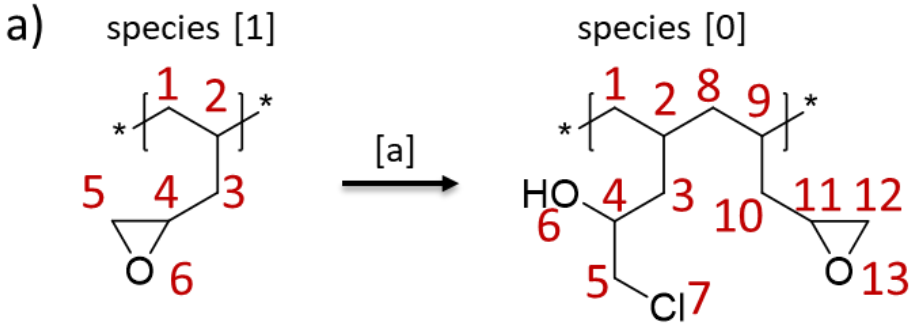
The *transformation* entry corresponds to an array of different transformations. Each transformation is denoted by a transformation-obj, which is composed of an ID (that matches that provided within the preamble) and a *contents* array composed of *component-obj* elements that further specify the content of the species:

```
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 tuples. 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, the oxygen atom in component [1:1] can be referenced using the label "[1:1]6". The syntax for the atom mapping is demonstrated in Figure S7.

Note that the provided generic schema for the transformation section is largely meant to serve as a basis for constructing schemas that are more specific to different chemical reactions and physicochemical processes. Therefore, only the pair of entries that are common to almost all transformations are specified within the base template. For examples that illustrate how augmentation of the base template could be applied to different processes, please refer to the following subsections.

a)



b)

```
"species": [
    {   "id" : "[1]",  "contents" : [
            {   "ID": "[1:1]",
                "bigsmiles" : "{[][$][C:1][C:2]([C:3][C:4]1[C:5][O:6]1)[$][]}"
    }   ]   },
    {   "id" : "[0]",  "contents" : [
                {   "ID" : "[0:1]",
                    "bigsmiles"  : "{[][$][C:1][C:2]([C:3][C:4]([C:5][Cl:7])[O:6])[$],
                                    [$][C:8][C:9]([C:10][C:11]1[C:12][O:13]1)[$][]}"
    }   ]   }
],
"transformation" : [
    {   "id" : "[a]",
        "atomMap" : [   ["[1:1]1", "[0:1]1", "[0:1]8" ], ["[1:1]2", "[0:1]2", "[0:1]9" ],
                        ["[1:1]3", "[0:1]3", "[0:1]10" ], ["[1:1]4", "[0:1]4", "[0:1]11" ],
                        ["[1:1]5", "[0:1]5", "[0:1]12"], ["[1:1]6", "[0:1]6", "[0:1]13" ]
        ]
    }
]
```

Figure S7. Illustration of the syntax and usage of the atom mapping scheme.

S5.1 Sample Schema for a Batch Reaction

Here, a template for specifying an isothermal, isobaric batch reaction is provided to demonstrate how the base template can be augmented to capture processes found in practical applications. Consider a transformation, [batch], with *n* feed species, [1], [2], …, [n], which includes both reactants, catalysts, and potentially solvents, and *one* output species, [o], which represents the mixture of the unreacted precursors and other converted species obtained after the reaction:

```
"network" : [
    "[1].[2].[3]. … .[n]>[batch]>[o]"
]
```

This batch reaction can be specified by the following sub-schema:

```
"transformation" : [
    {   "description" : "isobaric-isothermal-batch-reaction",
        "id" : string,
        "atomMap" : array of array of strings,
        "T" : number,
        "unit-T" : string,
        "P" : number,
        "unit-P" : string,
        "conversion" : [
            { "componentID" : string,  "substructure" : string, "conversion" : number },
            { "componentID" : string,  "substructure" : string, "conversion" : number },
            …
        ]
    }
]
```

in which the entries *T, P, unit-T* and *unit-P* are the temperature, pressure, and their corresponding units under which the batch reaction is carried out. Note that to fully determine the reaction condition, the initial condition of the system must be specified. Therefore, for each feed species, measurement of the amount of substance fed into the reactor must be provided explicit with the *quantity* entry. Finally, the overall conversion of the reaction, denoted by the conversions associated with individual reactants, are recorded.

For reactants that correspond to specific components within the list of feed species, the conversion can be simply specified by associating a numeric value with the *componentID*. For instance, the conversion for the third component within species [2] is encoded as:

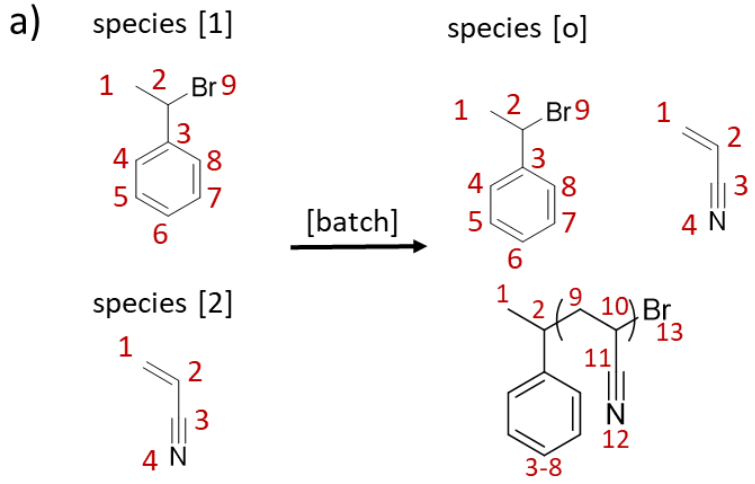
```
"conversion" : [
    { "componentID" : "[2:3]", "conversion" : 0.90 }, …
]
```

However, if conversion is defined to be the consumption of a specific functional group, the SMARTS pattern for the corresponding functional group should also be specified within the *conversion* array. For instance, to specify the consumption of an alkyne group, the following syntax can be used:

```
"conversion" : [
    { "componentID" : "[1:2]", "substructure" : "[C:1]#[C:2]", "conversion" : 0.90 }, …
]
```

Notably, it is not required to provide the conversion for all the species; only the reactants with their concentration monitored needs to be logged.

The usage of the provided sub-schema is demonstrated in Figure S8 through the polymerization of acrylonitrile. In this example, the reaction is carried out under 300K and 1bar (10^5 Pa), until a 90% conversion is reached.



b)

```
"species": [
    {   "id" : "[1]",  "contents" : [
            {   "ID":"[1:1]", "bigsmiles" : "[C:1][C:2]([c:3]1[c:4][c:5][c:6][c:7][c:8]1)[Br:9]",
                "quantity" : "0.040",   "units" : "UO_0000062"        }          ]          },
    {   "id" : "[2]",  "contents" : [
            {   "ID":"[2:1]", "bigsmiles" : "[C:1]=[C:2][C:3]#[C:4]" ]",
                "quantity" : "1.0",  "units" : "UO_0000062" }          ]          },
    {   "id" : "[o]",  "contents" : [
            {   "ID":"[o:1]", "bigsmiles" : "[C:1][C:2]([c:3]1[c:4][c:5][c:6][c:7][c:8]1)[Br:9]" },
            {   "ID":"[o:2]", "bigsmiles" : "[C:10]=[C:11][C:12]#[C:13]"    },
            {   "ID":"[o:p]",
                "bigsmiles" : "[C:14][C:15]([c:16]1[c:17][c:18][c:19][c:20][c:21]1)
                                {[$][$][C:22][C:23]([C:24]#[C:25][$][$]}[Br:26]" }
    ]   }
],
"transformation" : [
    {   "id" : "[batch]",
        "atomMap" : [   ["[1:1]1", "[o:1]1", "[o:p]1" ], ["[1:1]2", "[o:1]2", "[o:p]2" ],
                        ["[1:1]3", "[o:1]3", "[o:p]3" ], ["[1:1]4,8", "[o:1]4,8", "[o:p]4,8" ],
                        ["[1:1]5,7", "[o:1]5,7", "[o:p]5,7" ], ["[1:1]6", "[o:1]6", "[o:p]6" ],
                        ["[1:1]9", "[o:1]9" ],
                        ["[2:1]1", "[o:2]1", "[o:p]9" ], ["[2:1]2", "[o:2]2", "[o:p]10" ],
                        ["[2:1]3", "[o:2]3", "[o:p]11" ], ["[2:1]4", "[o:2]4", "[o:p]12" ]        ],
        "T" : 300,    "unit-T" : "UO_0000012",
        "P" : 100000,    "unit-P" : "UO_0000110",
        "conversion" : [
            {   "componentID" : "[2:1]",
                "substructure" : "[C:1]=[C:2][C:3]#[C:4]",
                "conversion": 0.9 }
        ]
    }
]
```

Figure S8. Illustrative example of a batch reaction encoded with the provided template. (a) Schematic illustration of the batch reaction of acrylonitrile using ATRP (ligand has been omitted for simplicity). (b) An implementation of the batch reaction template on the reaction specified in part (a).

S5.2 Sample Schema for a Fractionation Process

Similar to the batch reaction, a fractionation process can also be captured by augmenting the base template. Consider a fractionation process, [fract], with one feed species, [i], and *n* output species, [o1], [o2], …, [on], which represent the fractionated phases obtained after the equilibrium is attained upon a thermodynamic stimulation.

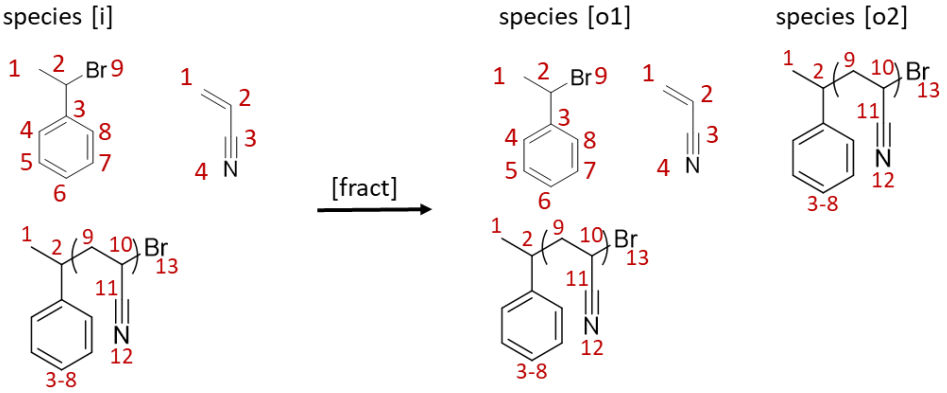
```
"network" : [
    "[i]>[fract]>[o1].[o2]. … .[on]"
]
```

This fractionation process can be specified by the following sub-schema:

```
"transformation" : [
    {   "description" : "equilibrium fractionation",
        "id" : string,
        "atomMap" : array of array of strings,
        "T" : number,
        "unit-T" : string,
        "P" : number,
        "unit-P" : string,
        "solvent" : array of component-obj,
    }
]
```

in which the entries *T, P, unit-T* and *unit-P* are the temperature, pressure, and their corresponding units under which the batch reaction is carried out, whereas *solvent* denotes the solvent associated with the process, specified by an array of *component-obj* collecting the SMILES strings and the concentrations for the component. Like the previous example for batch reaction, the extensive quantities for the contents within the feed, as well as the solvent, must be explicitly specified with the *quantity* field. Since the thermodynamic parameters fully determines the phase separation outcome of the equilibrium fractionation process, the template provides a compact form for encapsulating such process. Figure S9 provides a practical example on the usage of the proposed template.

a)



b)

```
"species" : [
    {   "ID" : "[i]",
        "contents" : [
            {   "quantity" : "10",    "units" : "g",
                "contents" : [
                    {   "ID":"[i:AN]", "bigsmiles" : "C=CC#N",
                        "quantity" : "0.60", "units" : "mol/L"     },
                    {   "ID":"[i:PEBr]", "bigsmiles" : "c1ccccc1C(C)Br",
                        "quantity" : "0.020", "units" : "mol/L"    },
                    {   "ID":"[i:polymer]", "bigsmiles" : "CC(c1ccccc1){[$][$]CC(C#N)[$][$]}Br",
                        "quantity" : "0.10", "units" : "mol/L"     },
                    {   "ID":"[i:solvent]", "bigsmiles" : "O(c1ccccc1)c2ccccc2"         }
                ]
            }
        ]
    }, { "ID" : "[o1]", ... }, { "ID" : "[o2]", ... }
],
"transformation" : [
    {   "id" : "[fract]",
        "atomMap" : [   … ],
        "T" : 270,
        "unit-T" : "UO_0000012",
        "P" : 100000,
        "unit-P" : "UO_0000110",
        "solvent" : [
            {   "bigsmiles" : "CN(C)C=O",
                "quantity" : "500",  "units" : "UO_0000021" }
        ]
    }
]
```

Figure S9. Illustrative example of a fractionation process encoded with the provided template. (a) Schematic illustration of the fractionation/precipitation of poly(acrylonitrile). (b) An implementation of the batch fractionation on the precipitation of poly(acrylonitrile) specified in part (a). Here, the residual polymer in [o1] reflects the case when precipitation does not completely isolate all polymer.

S6. Sample PolyDAT Files for Selected Polymers

In this section, the PolyDAT files for a few selected polymers are presented to further illustrate the syntax and utility of the schema.

A summary of the provided examples can be found in the following table:

Example	Page	Notes

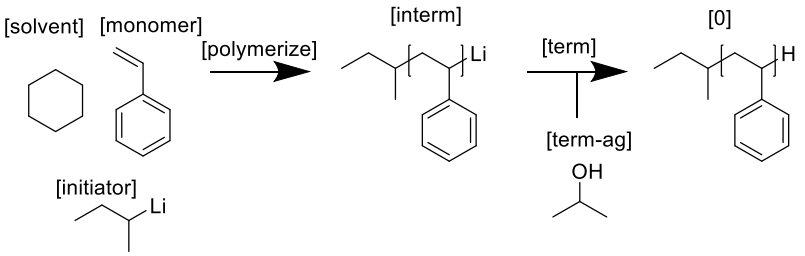
Polymer

Ionic polymerization of styrene.

Source

Lee, W., Lee, H., Cha, J., Chang, T., Hanley, K. J., & Lodge, T. P. (2000). Molecular weight distribution of polystyrene made by anionic polymerization. *Macromolecules*, 33(14), 5111-5115.

Schematics



Description & Notes

Anionic living polymerization of styrene.

PolyDAT


```
{
   "preamble": {
      "polymer": "CCC(C){[$][$]CC([$])c1ccccc1[$]}[H]",
      "pdVersion": 1, "mixfileVersion": 0.01, "docID": "PS-1",
      "logs": [
         {   "author": ["ORCID:https://orcid.org/0000-0002-7272-7140"],
             "date": "2020-07-01", "msg": "Created Entry" }      ],
      "srcs": [
         {   "citeID": "src1", "doi": "10.1016/0032-3861(95)99001-B",
             "desc": "Morphologies of strongly segregated polystyrene-poly(dimethylsiloxane) diblock
              copolymers; sample 9, table 1" }       ],
      "network": [ "[solvent].[monomer].[initiator]>[polymerize]>[interm]",
                   "[interm].[term-ag]>[term]>[0]"  ]
   },
   "species": [
      {   "ID": "[0]",
          "contents": [
             {   "ID": "[0:1]",
                 "bigsmiles": "CCC(C){[$][$][C:1][C:2]([$])[c:3]1[c:4][c:5][c:6][c:7][c:8]1[$]}[H]",
                 "characterization": {
                     "Mw": [ { "value": 4300, "unit": "UO_0000221", "src": "src1", "method": "gpc" } ],
                     "Mw/Mn": [ { "value": 1.02, "unit": "UO_0000186", "src": "src1", "method": "gpc" } ]
                 }
             }
          ]
      },
      {   "ID": "[interm]",
          "contents": [
             {   "ID": "[interm:1]",
```

```
                "bigsmiles": "CCC(C){[$][$][C:1][C:2]([$])[c:3]1[c:4][c:5][c:6][c:7][c:8]1[$]}[Li]"
            }
        ]
    },
    {   "ID": "[term-ag]",
        "contents": [
            {   "ID": "[term-ag:1]", "bigsmiles": "CC(O)C"  }
        ]
    },
    {   "ID": "[monomer]",
        "contents": [
            {   "ID": "[monomer:1]", "bigsmiles": "[C:1][C:2][c:3]1[c:4][c:5][c:6][c:7][c:8]1"     }
        ]
    },
    {   "ID": "[initiator]",
        "contents": [
            {   "ID": "[initiator:1]", "bigsmiles": "CCC(C)[Li]" }
        ]
    },
    {   "ID": "[solvent]",
        "contents": [
            {   "ID": "[solvent:1]", "bigsmiles": "C1CCCCC1" }
        ]
    }

  ],
  "transformation": [
    {   "ID": "[polymerize]",
        "atomMap": [
            [ "[monomer:1]1", "[interm:1]1" ],
            [ "[monomer:1]2", "[interm:1]2" ],
            [ "[monomer:1]3", "[interm:1]3" ],
            [ "[monomer:1]4,8", "[interm:1]4,8" ],
            [ "[monomer:1]5,7", "[interm:1]5,7" ],
            [ "[monomer:1]6", "[interm:1]6" ],
        ]
    },
    {   "ID": "[term]",
        "atomMap": [
            [ "[interm:1]1", "[0:1]1" ],
            [ "[interm:1]2", "[0:1]2" ],
            [ "[interm:1]3", "[0:1]3" ],
            [ "[interm:1]4,8", "[0:1]4,8" ],
            [ "[interm:1]5,7", "[0:1]5,7" ],
            [ "[interm:1]6", "[0:1]6" ]
        ]
    }
  ]
}
```

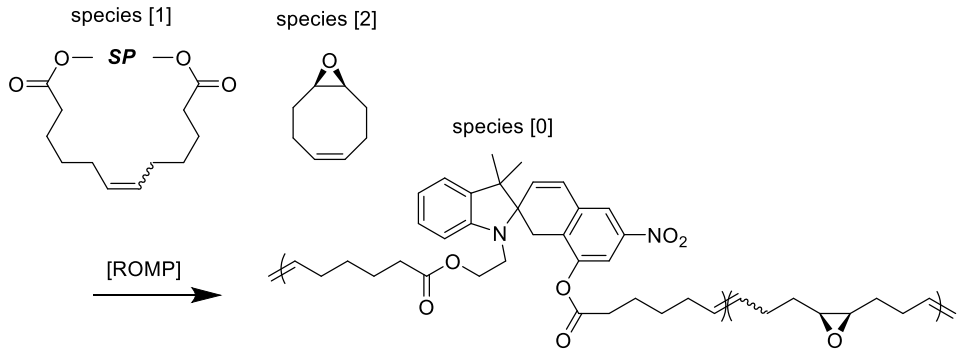
Polymer

Random copolymer with spiropyran.

Source

Gossweiler, G. R., Kouznetsova, T. B., & Craig, S. L. (2015). Force-rate characterization of two spiropyran-based molecular force probes. *Journal of the American Chemical Society*, *137*(19), 6148-6151.

Schematics



Description & Notes

In this example, only a few of selected atoms are labelled and included within the atom mapping section. This practice is permissible as long as the labelled atoms provide sufficient information such that the other atoms can be unambiguously mapped.

PolyDAT


```
{
   "preamble": {
      "polymer": "{[][$]=CCCCCC(=O)OCCN1c2ccccc2C(C)(C)C1Oc3c(cc1)
                        cc(N(O)O)cc3OC(=O)CCCCC=[$],[$]=CCCC1C(O1)CCC=[$][]}",
      "pdVersion": 1, "mixfileVersion": 0.01, "docID": "spiropyran-01",
      "logs": [
         {   "author_id": ["ORCID: https://orcid.org/0000-0002-7510-1194"],
             "date": "2020-09-06", "msg": "Created Entry" }      ],
      "srcs": [
         {   "citeID": "src1", "doi": "10.1021/jacs.5b02492",
             "desc": " Force-rate characterization of two spiropyran-based molecular force probes." }
      ],
      "network": [ "[1].[2]>[ROMP]>[0]" ]
   },
   "species": [
      {   "ID": "[1]",
          "contents": [
             {   "ID": "[1:1]",
                 "bigsmiles": "[C:1]=1[C:2]CCCC(=O)OCCN1[c:3]2[c:4][c:5][c:6][c:7][c:8]2C
                                 (C)(C)C1Oc3c(cc1)cc(N(O)O)cc3OC(=O)CCCC[C:41]1",
          }   ]
      },
```

```
        {   "ID": "[2]",
            "contents": [
                {   "ID": "[2:1]",
                    "bigsmiles": "[C:1]1[C:2][C:3][C:4]2[C:5]([O:6]2)[C:7][C:8][C:9]1"
            }   ]
        },
        {   "ID": "[0]",
            "contents": [
                {   "ID": "[0:1]",
                    "bigsmiles": ""{[][$]=[C:1][C:2]CCCC(=O)OCCN1[c:3]2[c:4][c:5][c:6][c:7][c:8]2
                                    C(C)(C)C1Oc3c(cc1)cc(N(O)O)cc3OC(=O)CCCC[C:41]=[$],[$]=[C:5
                                    1][C:52][C:53][C:54]1[C:55]([O:56]1)[C:57][C:58][C:59]=[$][]}",
                    "characterization": {
                        "ratios" : [
                            {
                                "substructure": [
                                    "[c:3][c:4][c:5][c:6][c:7][c:8]",
                                    "[C:54]1[C:55][O:56]1"
                                ],
                                "ratio" : [0.45,0.55],
                                "unit": "UO_0000013",
                                "method" : "nmr"
                            }
                    ]   }
            }   ]
        }
    ],
    "transformation": [
        {   "ID": "[ROMP]",
            "atomMap": [
                [ "[1:1]1", "[0:1]1" ],  [ "[1:1]2", "[0:1]2" ], [ "[1:1]3", "[0:1]3" ], [ "[1:1]4", "[0:1]4" ],
                [ "[1:1]5", "[0:1]5" ],  [ "[1:1]6", "[0:1]6" ], [ "[1:1]7", "[0:1]7" ], [ "[1:1]8", "[0:1]8" ],
                [ "[1:1]41", "[0:1]41" ],
                [ "[2:1]1,9", "[0:1]51,59" ],
                [ "[2:1]2,8", "[0:1]52,58" ],
                [ "[2:2]3,7", "[0:1]53,57" ],
                [ "[2:3]4,5", "[0:1]54,55" ],
                [ "[2:5]6", "[0:1]56" ]
            ]
        }
    ]
}
```

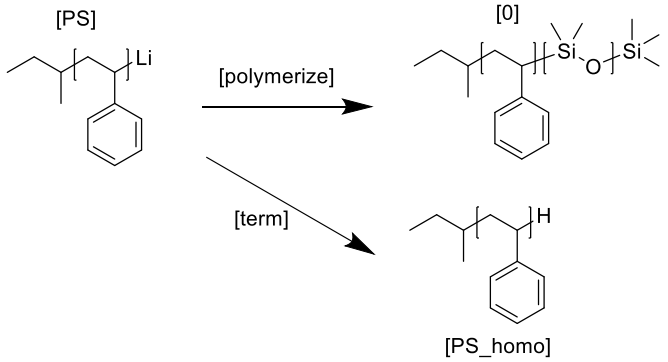
Polymer

PS-*b*-PDMS diblock copolymer

Source

Chu, J. H., Rangarajan, P., Adams, J. L., & Register, R. A. (1995). Morphologies of strongly segregated polystyrene-poly (dimethylsiloxane) diblock copolymers. *Polymer*, *36*(8), 1569-1575.

Schematics



Description & Notes

The polymerization of the PDMS block with PS precursor is encoded within the following file. This example illustrates how PolyDAT can be utilized to provide a simple way of encoding both information for the first block as well as the entire block copolymer.

PolyDAT


```
{
    "preamble": {
        "polymer": "CCC(C){[$][$]CC([$])c1ccccc1[$]}{[>][>]O[Si]([<])(C)C[<]}[Si](C)(C)C ",
        "pdVersion": 1, "mixfileVersion": 0.01, "docID": "PS-PDMS-1",
        "logs": [
            {   "author": ["ORCID:https://orcid.org/0000-0002-7272-7140"],
                "date": "2020-07-01", "msg": "Created Entry" }      ],
        "srcs": [
            {   "citeID": "src1", "doi": "10.1016/0032-3861(95)99001-B",
                "desc": "Morphologies of strongly segregated polystyrene-poly(dimethylsiloxane) diblock
                 copolymers; sample 9, table 1" }       ],
        "network": [ "[PS]>[polymerize]>[0]", "[PS].[MeOH]>[term]>[PS_homo]"  ]
    },

    "species": [
        {   "ID": "[0]",
            "contents": [
                {   "ID": "[0:1]",
                    "bigsmiles": "CCC(C){[$][$][C:1][C:2]([$])[c:3]1[c:4][c:5][c:6][c:7][c:8]1[$]}
                                  {[>][>][O:9][Si:10]([<])([C:11])[C:12][<]}[Si](C)(C)C",
                    "characterization": {
                        "Mw": [ { "value": 68100, "unit": "UO_0000221", "src": "src1", "method": "gpc" } ],
                        "Mw/Mn": [ { "value": 1.12, "unit": "UO_0000186", "src": "src1", "method": "gpc" } ]
                    }
                }
```

```
            ]
        },
        {   "ID": "[PS]",
            "contents": [
                {   "ID": "[PS:1]",
                    "bigsmiles": "CCC(C){[$][$][C:1][C:2]([$])[c:3]1[c:4][c:5][c:6][c:7][c:8]1[$]}[Li]"
                }
            ]
        },
        {   "ID": "[PS_homo]",
            "contents": [
                {   "ID": "[PS_homo:1]",
                    "bigsmiles": "CCC(C){[$][$][C:1][C:2]([$])[c:3]1[c:4][c:5][c:6][c:7][c:8]1[$]}[H]",
                    "characterization": {
                        "Mw": [ { "value": 5700, "unit": "UO_0000221", "src": "src1", "method": "gpc" } ],
                        "Mw/Mn": [ { "value": 1.08, "unit": "UO_0000186", "src": "src1", "method": "gpc" } ]
                    }
                }
            ]
        },
        {   "ID": "[MeOH]",
            "contents": [
                {   "ID": "[MeOH:1]", "bigsmiles": "CO" }
            ]
        }
    ],
    "transformation": [
        {   "ID": "[polymerize]",
            "atomMap": [
                [ "[PS:1]1", "[0:1]1" ],
                [ "[PS:1]2", "[0:1]2" ],
                [ "[PS:1]3", "[0:1]3" ],
                [ "[PS:1]4,8", "[0:1]4,8" ],
                [ "[PS:1]5,7", "[0:1]5,7" ],
                [ "[PS:1]6", "[0:1]6" ],
            ]
        },
        {   "ID": "[term]",
            "atomMap": [
                [ "[PS:1]1", "[PS_homo:1]1" ],
                [ "[PS:1]2", "[PS_homo:1]2" ],
                [ "[PS:1]3", "[PS_homo:1]3" ],
                [ "[PS:1]4,8", "[PS_homo:1]4,8" ],
                [ "[PS:1]5,7", "[PS_homo:1]5,7" ],
                [ "[PS:1]6", "[PS_homo:1]6" ]
            ]
        }
    ]
}
```

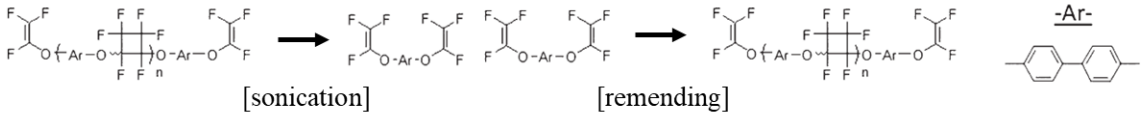
Polymer

Mechanically induced scission and thermal remending of PFCB polymer.

Source

Klukovich, H. M., Kean, Z. S., Iacono, S. T., & Craig, S. L. (2011). Mechanically induced scission and subsequent thermal remending of perfluorocyclobutane polymers. *Journal of the American Chemical Society*, *133*(44), 17882-17888.

Schematics



Description & Notes

In this example, PolyDAT is used to capture a series of polymers samples that underwent sequential stress-induced chain scission and thermal re-mending. This example showcases the versatility of experimental procedures that can be encoded using the schema.

PolyDAT


```
{
   "preamble": {
      "polymer": "FC(F)=C(F)O{[>][<]c1ccc(cc1)c2ccc(cc2)C3(F)C(F)(F)C(F)(F)C(F)3[>][<]}O
                 c1ccc(cc1)c2ccc(cc2)OC(F)=C(F)F",
      "pdVersion": 1, "mixfileVersion": 0.01, "docID": "PFCB-01",
      "logs": [
         {  "author_id": ["ORCID: https://orcid.org/0000-0002-7510-1194"],
            "date": "2020-09-06", "msg": "Created Entry" }     ],
      "srcs": [
         {  "citeID": "src1", "doi": "10.1021/ja2074517",
            "desc": "Mechanically Induced Scission and Subsequent Thermal Remending of
                     Perfluorocyclobutane Polymers" }
      ],
      "network": [ "[1]>[sonication1]>[2]",
                   "[2]>[remending1]>[3]",
                   "[3]>[sonication2]>[4]",
                   "[4]>[remending2]>[0]" ]
   },
   "species": [
      {  "ID": "[1]",
         "contents": [
            {  "ID": "[1:1]",
               "bigsmiles": "FC(F)=C(F)[O:1]{[>][<]c1ccc(cc1)c2ccc(cc2)C3(F)C(F)(F)C(F)(F)C(F)3
                            [>][<]}[O:2]c1ccc(cc1)c2ccc(cc2)OC(F)=C(F)F",
               "characterization": {
                  "Mn": [ { "value":115, "unit":"UO_0000222", "src":"src1", "method":"gpc" } ]
               }
         }     ]
      },
```

```
        {   "ID": "[2]",
            "contents": [
                {   "ID": "[2:1]",
                    "bigsmiles": "FC(F)=C(F)[O:1]{[>][<]c1ccc(cc1)c2ccc(cc2)C3(F)C(F)(F)C(F)(F)C(F)3
                                 [>][<]}[O:2]c1ccc(cc1)c2ccc(cc2)OC(F)=C(F)F",
                    "characterization": {
                        "Mn": [ { "value":10, "unit":"UO_0000222", "src":"src1", "method":"gpc" } ]
                    }
        }       ]       },
        {   "ID": "[3]",
            "contents": [
                {   "ID": "[3:1]",
                    "bigsmiles": "FC(F)=C(F)[O:1]{[>][<]c1ccc(cc1)c2ccc(cc2)C3(F)C(F)(F)C(F)(F)C(F)3
                                 [>][<]}[O:2]c1ccc(cc1)c2ccc(cc2)OC(F)=C(F)F",
                    "characterization": {
                        "Mn": [ { "value":37, "unit":"UO_0000222", "src":"src1", "method":"gpc" } ]
                    }
        }       ]       },
        {   "ID": "[4]",
            "contents": [
                {   "ID": "[4:1]",
                    "bigsmiles": "FC(F)=C(F)[O:1]{[>][<]c1ccc(cc1)c2ccc(cc2)C3(F)C(F)(F)C(F)(F)C(F)3
                                 [>][<]}[O:2]c1ccc(cc1)c2ccc(cc2)OC(F)=C(F)F",
                    "characterization": {
                        "Mn": [ { "value":13, "unit":"UO_0000222", "src":"src1", "method":"gpc" } ]
                    }
        }       ]       },
        {   "ID": "[0]",
            "contents": [
                {   "ID": "[0:1]",
                    "bigsmiles": "FC(F)=C(F)[O:1]{[>][<]c1ccc(cc1)c2ccc(cc2)C3(F)C(F)(F)C(F)(F)C(F)3
                                 [>][<]}[O:2]c1ccc(cc1)c2ccc(cc2)OC(F)=C(F)F",
                    "characterization": {
                        "Mn": [ { "value":36, "unit":"UO_0000222", "src":"src1", "method":"gpc" } ]
                    }
        }       ]       }
    ],
    "transformation": [
        {   "ID": "[sonication1]",
            "atomMap": [ [ "[1:1]1", "[2:1]1" ], [ "[1:1]2", "[2:1]2" ] ]  },
        {   "ID": "[remending1]",
            "atomMap": [ [ "[2:1]1", "[3:1]1" ], [ "[2:1]2", "[3:1]2" ] ]  },
        {   "ID": "[sonication2]",
            "atomMap": [ [ "[3:1]1", "[4:1]1" ], [ "[3:1]2", "[4:1]2" ] ]  },
        {   "ID": "[remending2]",
            "atomMap": [ [ "[4:1]1", "[0:1]1" ], [ "[4:1]2", "[0:1]2" ] ]  },
    ]
}
```

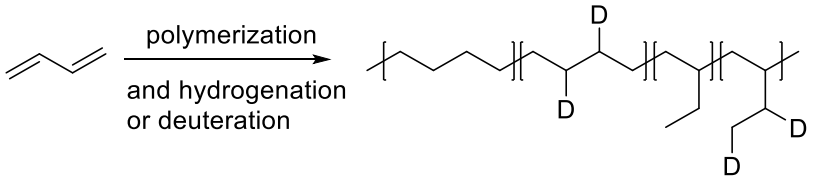
Polymer

Partially deuterated polyolefin.

Source

Balsara, N. P., Fetters, L. J., Hadjichristidis, N., Lohse, D. J., Han, C. C., Graessley, W. W., & Krishnamoorti, R. (1992). Thermodynamic interactions in model polyolefin blends obtained by small-angle neutron scattering. *Macromolecules*, *25*(23), 6137-6147.

Schematics



Description & Notes

Because of the symmetry around the terminal carbons and the central carbons for the precursor diene, the atom mappings for the reaction are many-to-many.

PolyDAT


```
{
    "preamble": {
        "polymer": "{[][$]CCCC[$],[$]CC(D)C(D)C[$],[$]CC(CC)[$],[$]CC(C(D)C(D))[$][]}",
        "pdVersion": 1, "mixfileVersion": 0.01, "docID": "deuterated-01",
        "logs": [
            {   "author_id": ["ORCID: https://orcid.org/0000-0002-7510-1194"],
                "date": "2020-09-06", "msg": "Created Entry" }      ],
        "srcs": [
            {   "citeID": "src1", "doi": "10.1021/ma00049a009",
                "desc": "Thermodynamic interactions in model polyolefin blends obtained by small-angle
                        neutron scattering." }
        ],
        "network": [ "[1]>[polymerize]>[0]" ]
    },
    "species": [
        {   "ID": "[1]",
            "contents": [
                {   "ID": "[1:1]",
                    "bigsmiles": "[C:1]=[C:2][C:3]=[C:4]",
                },
                {   "ID": "[1:2]",
                    "bigsmiles": "[H][H]",
                },
                {   "ID": "[1:3]",
                    "bigsmiles": "[2H][2H]",
                }
            ]
        },
        {   "ID": "[0]",
```

```
            "contents": [
                {   "ID": "[0:1]",
                    "bigsmiles": "{[][$][C:11][C:12][C:13][C:14][$],[$][C:21][C:22](D)[C:23](D)[C:24][$],
                                  [$][C:31][C:32]([C:33][C:34])[$],[$][C:41][C:42]([C:43](D)[C:44](D))[$]
                                  []}",
                    "characterization": {
                        "DPn": [ { "value":36, "unit":"UO_0000186", "src":"src1", "method":"scattering" } ],
                        "Mw/Mn": [ { "value":1.06, "unit":"UO_0000186", "src":"src1" } ],
                        "ratios" : [
                            {
                                "substructure": [ "[1H]", "[2H]" ],
                                "ratio" : [0.35,0.65],
                                "unit": "UO_0000013"
                            }
                        ]
                    }
                }
            ]
        }
    ],
    "transformation": [
        {   "ID": "[polymerize]",
            "atomMap": [
                [   "[1:1]1,4", "[0:1]11,14", "[0:1]21,24", "[0:1]31", "[0:1]34", "[0:1]41", "[0:1]44" ],
                [   "[1:1]2,3", "[0:1]12,13", "[0:1]22,23", "[0:1]32", "[0:1]33", "[0:1]42", "[0:1]43" ]
            ]
        }
    ]
}
```

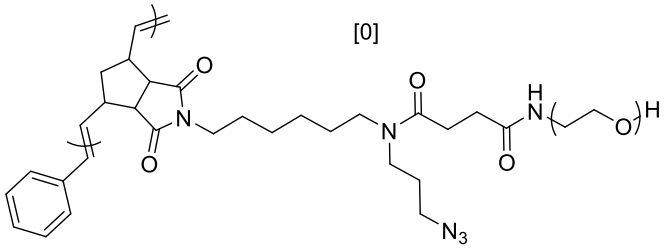
Polymer

Bottle-brush polymer.

Source

Johnson, J. A. et al. (2011). Core-clickable PEG-branch-azide bivalent-bottle-brush polymers by ROMP: grafting-through and clicking-to. *Journal of the American Chemical Society*, *133*(3), 559-566.

Schematics



Description & Notes

In this example, PolyDAT's capability to encapsulate an entire experimental synthetic procedure is demonstrated by encoding the synthetic procedure of a bottle-brush polymer.

PolyDAT


```
{
    "preamble": {
        "polymer": "C={[$][$]=CC1CC(C=[$])C2C(=O)N(CCCCCCN(CCCCl)C(=O)CCC(=O)N
                        {[>][<]CCO[>][<]}[H])C(=O)C12[$]}=Cc1ccccc1",
        "pdVersion": 1, "mixfileVersion": 0.01, "docID": "bottlebrush-01",
        "logs": [
            {   "author_id": ["ORCID: https://orcid.org/0000-0002-7510-1194"],
                "date": "2020-09-06", "msg": "Created Entry" }      ],
        "srcs": [
            {   "citeID": "src1", "doi": "10.1021/ja108441d",
                "desc": "Core-Clickable PEG-Branch-Azide Bivalent-Bottle-Brush Polymers by ROMP." }
        ],
        "network": [
            "[species1].[species2]>[transformationA]>[species3]",
            "[species3].[species4].[species5]>[transformationB]>[species6]",
            "[species6].[species7].[species8].[species9]>[transformationC]>[species10]",
            "[species10].[species11]>[transformationD>[species12]",
            "[species12].[species13]>[transformationE]>[species14]",
            "[species14].[species15]>[transformationF]>[species16]",
            "[species16]>[transformationG]>[species17]",
            "[species17]>[transformationH]>[0]"
        ]
    },

    "species": [
        {   "ID": "[species1]",
            "contents": [
                {   "ID": "[species1:1]",
                    "bigsmiles": "O=C1OC(=O)[C@@H]2[C@@H]3C[C@@H](C=C3)[C@H]12",
```

```
            "quantity": 4, "units": "UO_0000021"
        }
    ]
},
{   "ID": "[species2]",
    "contents": [
        {   "ID": "[species2:1]",
            "bigsmiles": "NCCCCCCO",
            "quantity": 3, "units": "UO_0000021"
        }
    ]
},
{   "ID": "[species3]",
    "contents": [
        {   "ID": "[species3:1]",
            "bigsmiles": "O=C1N(CCCCCCO)C(C2C3C=CC(C3)C21)=O",
            "quantity": 6, "units": "UO_0000021"
        }
    ]
},
{   "ID": "[species4]",
    "contents": [
        {   "ID": "[species4:1]",
            "bigsmiles": "ClC(C(Cl)=O)=O",
            "quantity": 3.21, "units": "UO_0000098"
        }
    ]
},
{   "ID": "[species5]",
    "contents": [
        {   "ID": "[species5:1]",
            "bigsmiles": "CCN(CC)CC",
            "quantity": 20.83, "units": "UO_0000098"
        }
    ]
},
{   "ID": "[species6]",
    "contents": [
        {   "ID": "[species6:1]",
            "bigsmiles": "O=C1N(CCCCCC=O)C(C2C3C=CC(C3)C21)=O",
            "quantity": 5.83, "units": "UO_0000021"
        }
    ]
},
{   "ID": "[species7]",
    "contents": [
        {   "ID": "[species7:1]",
            "bigsmiles": "Cl[H].NCCCCl",
            "quantity": 523, "units": "UO_0000022"
        }
    ]
```

```
},
{   "ID": "[species8]",
    "contents": [
        {   "ID": "[species8:1]",
            "bigsmiles": "C(=O)([O-])[O-].[Na+].[Na+]",
            "quantity": 426, "units": "UO_0000022"
        }
    ]
},
{   "ID": "[species9]",
    "contents": [
        {   "ID": "[species9:1]",
            "bigsmiles": "[BH4-].[Na+]",
            "quantity": 232, "units": "UO_0000022"
        }
    ]
},
{   "ID": "[species10]",
    "contents": [
        {   "ID": "[species10:1]",
            "bigsmiles": "O=C1N(CCCCCCNCCCCl)C(C2C3C=CC(C3)C21)=O",
            "quantity": 1.1, "units": "UO_0000021"
        }
    ]
},
{   "ID": "[species11]",
    "contents": [
        {   "ID": "[species11:1]",
            "bigsmiles": "O=C1OC(=O)CC1",
            "quantity": 134, "units": "UO_0000022"
        }
    ]
},
{   "ID": "[species12]",
    "contents": [
        {   "ID": "[species12:1]",
            "bigsmiles": "O=C1N(CCCCCCN(C(CC(O)=O)=O)CCCCl)C(C2C3C=CC(C3)C21)=O",
            "quantity": 394, "units": "UO_0000022"
        }
    ]
},
{   "ID": "[species13]",
    "contents": [
        {   "ID": "[species13:1]",
            "bigsmiles": "O=C1N(O)C(=O)CC1",
            "quantity": 157, "units": "UO_0000022"
        }
    ]
},
{   "ID": "[species14]",
    "contents": [
```

```
            {   "ID": "[species14:1]",
                "bigsmiles": "O=C1N(CCCCCCN(C(CC(ON2C(CCC2=O)=O)=O)=O)CCCCl)
                             C(C3C4C=CC(C4)C31)=O",
                "quantity": 391, "units": "UO_0000022"
            }
        ]
    },
    {   "ID": "[species15]",
        "contents": [
            {   "ID": "[species15:1]",
                "bigsmiles": "[H]{[>][<]CCO[>][<]}N",
                "quantity": 134, "units": "UO_0000022",
                "characterization": { "Mn": [ { "value":3000, "unit":"UO_0000221", "src":"src1" } ] }
            }
        ]
    },
    {   "ID": "[species16]",
        "contents": [
            {   "ID": "[species16:1]",
                "bigsmiles": "[H]{[>][<]CCO[>][<]}NC(=O)CCC(=O)N(CCCCl)CCCCCCN3C(=C)
                             C2C1/C=C\\C(C1)C2C3=C",
                "quantity": 112, "units": "UO_0000022",
            }
        ]
    },
    {   "ID": "[species17]",
        "contents": [
            {   "ID": "[species17:1]",
                "bigsmiles": "C={[$][$]=CC1CC(C=[$])C2C(=O)N(CCCCCCN(CCCCl)C(=O)CCC
                             (=O)N{[>][<]CCO[>][<]}[H])C(=O)C12[$]}=Cc1ccccc1"
            }
        ]
    },
    {   "ID": "[0]",
        "contents": [
            {   "ID": "[0:1]",
                "bigsmiles": "C={[$][$]=CC1CC(C=[$])C2C(=O)N(CCCCCCN(CCCN=[N+]=[N-])
                             C(=O)CCC(=O)N{[>][<]CCO[>][<]}[H])C(=O)C12[$]}=Cc1ccccc1",
                "characterization": {
                    "Mn": [ { "value":343000, "unit":"UO_0000221", "src":"src1", "method":"gpc" } ],
                    "DPn": [ { "value":101, "unit":"UO_0000186", "src":"src1" } ],
                    "Mw/Mn": [ { "value":1.06, "unit":"UO_0000186", "src":"src1", "method":"gpc" } ],
                }
            }
        ]
    }
}
"transformation": [ ]
}
```

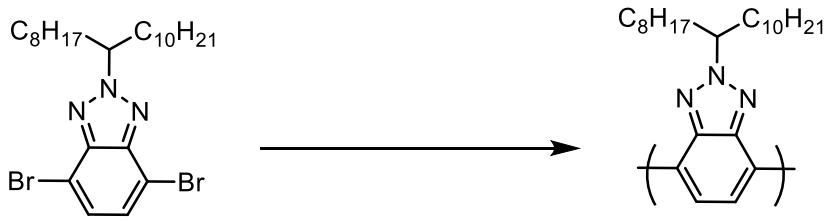
Polymer

Conjugate polymer.

Source

Woods, E. F., Berl, A. J., & Kalow, J. A. (2020). Photocontrolled synthesis of n-type conjugated polymers. *Angewandte Chemie, 132*(15), 6118-6123.

Schematics



Description & Notes

PolyDAT


```
{
    "preamble": {
        "polymer": "Br{[$][$]c1ccc(c2nn(C(CCCCCCCC)CCCCCCCCCC)nc12)[$][$]}Br",
        "pdVersion": 1, "mixfileVersion": 0.01, "docID": "conjugate-01",
        "logs": [
            {   "author_id": ["ORCID: https://orcid.org/0000-0002-7510-1194"],
                "date": "2020-09-06", "msg": "Created Entry" }      ],
        "srcs": [
            {   "citeID": "kalow", "doi": "10.1002/ange.201915819",
                "desc": "Photocontrolled Synthesis of n-Type Conjugated Polymers." }
        ],
        "network": [ "[precursor]>[a]>[intermediate]", "[intermediate].[salt]>[b]>[0]" ]
    },

    "species": [
        {   "ID": "[precursor]",
            "contents": [
                {   "ID": "[precursor:1]",
                    "bigsmiles": "Brc1ccc(c2nn(C(CCCCCCCC)CCCCCCCCCC)nc12)Br",
                    "name": "2,5-dibromobenzatriazole",
                    "quantity": 0.045, "unit": "UO_0000013"
                },
                {   "ID": "[precursor:2]",
                    "bigsmiles": "LiCCCC",
                    "name": "n-butyl lithium",
                    "quantity": 0.0405, "unit": "UO_0000013"
                },
                {   "ID": "[precursor:3]",
                    "bigsmiles": "([Mg].[Br]).CCOCC",
                    "name": "MgBr2et2O",
                    "quantity": 0.045, "unit": "UO_0000013"
```

```
            }
        ]
    },
    {   "ID": "[intermediate]",
        "contents": [
            {   "ID": "[intermediate:1]",
                "bigsmiles": "([Mg].[Br])Brc1ccc(c2nn(C(CCCCCCCC)CCCCCCCCCC)nc12)Br"
            }
        ]
    },
    {   "ID": "[salt]",
        "contents": [
            {   "ID": "[salt:1]",
                "bigsmiles": "[Li]Cl",
                "quantity": 0.675, "unit": "UO_0000013"
            }
        ]
    },
    {   "ID": "[0]",
        "contents": [
            {   "ID": "[salt:1]",
                "bigsmiles": "Br{[$][$]c1ccc(c2nn(C(CCCCCCCC)CCCCCCCCCC)nc12)[$][$]}Br",
                "quantity": 0.675, "unit": "UO_0000013",
                "characterization": {
                    "Mn": [ { "value":13.7, "unit":"UO_0000222", "src":"kalow", "method":"gpc" } ],
                    "Mw/Mn": [ { "value":1.61, "unit":"UO_0000186", "src":" kalow", "method":"gpc" } ]
                }
            }
        ]
    }
],
"transformation": [
    {   "ID": "[a]",
        "atomMap": [    ],
        "description": "isobaric-isothermal-batch-reaction"
        "T": 195, "unit-T": "UO_0000012",
        "P": 100000, "unit-P": "UO_0000110",
        "solvent": [ { "bigsmiles" : "C1CCOC1"} ]
    },
    {   "ID": "[b]",
        "atomMap": [    ],
        "description": "isobaric-isothermal-batch-reaction"
        "T": 296, "unit-T": "UO_0000012",
        "P": 100000, "unit-P": "UO_0000110",
        "solvent": [ { "bigsmiles" : "C1CCOC1"} ]
        "conversion": [ { "componentID" : "[intermediate:1]", "conversion": 0.9 } ]
    }
]
}
```

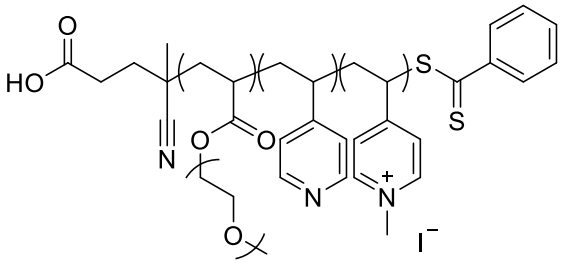
Polymer

POEGMA-*b*-qp4VP

Source

Obermeyer, A. C., Mills, C. E., Dong, X. H., Flores, R. J., & Olsen, B. D. (2016). Complex coacervation of supercharged proteins with polyelectrolytes. *Soft Matter*, *12*(15), 3570-3581.

Schematics



PolyDAT


```
{
    "preamble": {
        "polymer": "CC(C#N)(CCC(=O)O){[$][$]CC([$])(C)C(=O)O{[>][<]CCO[>][<]}C[$]}
                    {[$][$]CC([$])c1cc[n+](C)cc1.[I-],[$]CC([$])c1ccncc1[$]}SC(=S)c1ccccc1",
        "pdVersion": 1, "mixfileVersion": 0.01, "docID": "POEGMA-b-q4VP-1",
        "logs": [
            {   "author": ["ORCID:https://orcid.org/0000-0002-7272-7140"],
                "date": "2020-07-01", "msg": "Created Entry" }      ],
        "srcs": [
            {   "citeID": "src1", "doi": "10.1039/c6sm00002a ",
                "desc": "" }            ],
        "network": [ "[1]>[a]>[2]",
                     "[2].[3]>[b]>[4]",
                     "[4]>[c]>[0]"  ]
    },
    "species": [
        {   "ID": "[0]",
            "contents": [
                {   "ID": "[0:1]",
                    "bigsmiles": "[CH3:a12][C:a5]([C:a11]#[N:a13])([CH2:a4][CH2:a3][C:a2](=[O:a10])
                                  [OH:a1]){[$][$][CH2:6R1a2][C:6R1a3]([$])([CH3:6R1a6])[C:6R1a5](=[O:
                                  6R1a10])[O:6R1a7]{[>][<][CH2:6R1a8R1a4][CH2:6R1a8R1a3][O:6R1a8
                                  R1a2][>][<]}[CH3:6R1a9][$]}{[$][$][CH2:7R1a2][CH:7R1a3]([$])[c:7R1
                                  a5]1[cH:7R1a10][cH:7R1a9][n+:7R1a8]([CH3:7R1a11])[cH:7R1a7][cH:7
                                  R1a6]1.[I-:7R1a12],[$][CH2:7R2a2][CH:7R2a3]([$])[c:7R2a5]1[cH:7R2a1
                                  0][cH:7R2a9][n:7R2a8][cH:7R2a7][cH:7R2a6]1
[$]}[S:a8][C:a9](=[S:a15])[c:a14]1[cH:a20][cH:a19][cH:a18][cH:a17][cH:a16]1"
                }
            ]
        },
        {   "ID": "[1]",
            "contents": [
```

```
        {   "ID": "[1:1]",
            "bigsmiles": "[CH2:a1]=[C:a2]([CH3:a4])[C:a3](=[O:a7])[O:a5]{[<][>][CH2:6R1a4]
                         [CH2:6R1a3][O:6R1a2][<][>]}[CH3:a8]",
            "quantity": 100, "unit": "UO_0000040",
            "characterization": {
                "Mn": [ { "value": 300, "unit": "UO_0000221", "src": "src1" } ]
            }
        },
        {   "ID": "[1:2]",
            "bigsmiles": "[CH3:a14][C:a9]([C:a15]#[N:a16])([CH2:a10][CH2:a11][C:a12]
                         (=[O:a18])[OH:a17])[S:a8][C:a7](=[S:a13])[c:a2]1[cH:a3][cH:a4][cH:a5][c
                         H:a6][cH:a1]1",
            "quantity": 0.56, "unit": "UO_0000040"
        },
        {   "ID": "[1:3]",
            "bigsmiles": "[CH3:a10][C:a5]([CH3:a6])([C:a9]#[N:a12])/[N:a4]=[N:a3]/[C:a2]
                         ([CH3:a8])([CH3:a1])[C:a7]#[N:a11]",
            "quantity": 0.11, "unit": "UO_0000040"
        },
        {   "ID": "[1:4]",
            "bigsmiles": "[CH2:a4]1[CH2:a5][O:a6][CH2:a1][CH2:a2][O:a3]1"
        }
    ]
},
{   "ID": "[2]",
    "contents": [
        {   "ID": "[2:1]",
            "bigsmiles": "[CH3:a11][C:a5]([C:a10]#[N:a12])([CH2:a4][CH2:a3][C:a2](=[O:a9])
                         [OH:a1]){[$][$][CH2:6R1a2][C:6R1a3]([$])([CH3:6R1a6])[C:6R1a5](=[O
                         :6R1a10])[O:6R1a7]{[>][<][CH2:6R1a8R1a2][CH2:6R1a8R1a3][O:6R1a8
                         R1a4][>][<]}[CH3:6R1a9][$]}[S:a7][C:a8](=[S:a19])[c:a13]1[cH:a18][cH:
                         a17][cH:a16][cH:a15][cH:a14]1"
        }
    ]
},
{   "ID": "[3]",
    "contents": [
        {   "ID": "[3:1]",
            "bigsmiles": "[CH2:a1]=[CH:a2][c:a3]1[cH:a8][cH:a7][n:a6][cH:a5][cH:a4]1",
            "quantity": 23.1, "unit": "UO_0000040"
        },
        {   "ID": "[3:2]",
            "bigsmiles": "[CH3:a9][C:a5]([CH3:a6])([C:a10]#[N:a12])/[N:a4]=[N:a3]/[C:a2]
                         ([CH3:a8])([CH3:a1])[C:a7]#[N:a11]",
            "quantity":0.01, "unit": "UO_0000040"
        },
        {   "ID": "[3:3]",
            "bigsmiles": " C1COCCO1"
        },
        {   "ID": "[3:4]",
            "bigsmiles": " CN(C)C=O",
```

```
            }
        ]
    },
    {   "ID": "[4]",
        "contents": [
            {   "ID": "[4:1]",
                "bigsmiles": "[CH3:a12][C:a6]([C:a11]#[N:a13])([CH2:a5][CH2:a2][C:a1](=[O:a3])
                                [OH:a4]){[$][$][CH2:7R1a2][C:7R1a3]([$])([CH3:7R1a6])[C:7R1a5](=[O
                                :7R1a9])[O:7R1a7]{[>][<][CH2:7R1a8R1a4][CH2:7R1a8R1a3][O:7R1a8R
                                1a2][>][<]}[CH3:7R1a10][$]}{[$][$][CH2:8R1a1][CH:8R1a3]([$])[c:8R1a
                                5]1[cH:8R1a10][cH:8R1a9][n:8R1a8][cH:8R1a7][cH:8R1a6]1[$]}[S:a9][C
                                :a10](=[S:a15])[c:a14]1[cH:a20][cH:a19][cH:a18][cH:a17][cH:a16]1",
                "characterization": {
                    "Mn": [ { "value": 48.5, "unit": "UO_0000222", "src": "src1", "method": "gpc" } ],
                    "Mw/Mn": [ { "value": 1.13, "unit": "UO_0000186", "src": "src1", "method": "gpc" } ]
                }
            },
            {   "ID": "[4:2]",
                "bigsmiles": "[CH3:a2][I:a1]"
            },
            {   "ID": "[4:3]",
                "bigsmiles": " CN(C)C=O",
            }
        ]
    }
],
"transformation": [
    {   "ID": "[a]",
        "atomMap": [
            [ "[1:2]a17", "[2:1]a1" ], ["[1:2]a12", "[2:1]a2" ], [ "[1:2]a18", "[2:1]a9" ],
            [ "[1:2]a11", "[2:1]a3" ], ["[1:2]a10", "[2:1]a4" ], [ "[1:2]a9", "[2:1]a5" ],
            [ "[1:2]a14", "[2:1]a11"], ["[1:2]a15", "[2:1]a10" ], [ "[1:2]a16", "[2:1]a12"],
            [ "[1:2]a8", "[2:1]a7"], [ "[1:2]a7", "[2:1]a8" ], [ "[1:2]a13", "[2:1]a19" ],
            [ "[1:2]a2", "[2:1]a13" ], [ "[1:2]a1,a3","[2:1]a14,a18"],
            [ "[1:2]a5", "[2:1]16" ], [ "[1:2]a4,6","[2:1]a15,a17"],
            ["[1:1]a1","[2:1]6R1a2"], ["[1:1]a2","[2:1]6R1a3"], ["[1:1]a4","[2:1]6R1a6"],
            ["[1:1]a3","[2:1]6R1a5"], ["[1:1]a7","[2:1]6R1a10"], ["[1:1]a5","[2:1]6R1a7"],
            ["[1:1]6R1a4","[2:1]6R1a8R1a2"], ["[1:1]6R1a3","[2:1]6R1a8R1a3"],
            ["[1:1]6R1a2","[2:1]6R1a8R1a4"], ["[1:1]a8","[2:1]6R1a9"]
        ],
    },
    {   "ID": "[b]",
        "atomMap": [
            ["[2:1]a1","[4:1]a4"], ["[2:1]a2","[4:1]a2"], ["[2:1]a9","[4:1]a3"], ["[2:1]a3","[4:1]a2"],
            ["[2:1]a4","[4:1]a5"], ["[2:1]a5","[4:1]a6"], ["[2:1]a11","[4:1]a12"],
            ["[2:1]a10","[4:1]a11"],   ["[2:1]a12","[4:1]a13"], ["[2:1]6R1a2","[4:1]6R1a2"],
            ["[2:1]6R1a3","[4:1]6R1a3"], ["[2:1]6R1a6","[4:1]6R1a6"],["[2:1]6R1a5","[4:1]6R1a5"],
            ["[2:1]6R1a10","[4:1]6R1a9"],  ["[2:1]6R1a7","[4:1]6R1a7"],
            ["[2:1]6R1a9","[4:1]6R1a10"], ["[2:1]6R1a8R1a2","[4:1]6R1a8R1a4"],
            ["[2:1]6R1a8R1a3","[4:1]6R1a8R1a3"],   ["[2:1]6R1a8R1a4","[4:1]6R1a8R1a2"],
```

```
            ["[2:1]a7","[4:1]a9"], ["[2:1]a8","[4:1]a10"],["[2:1]a19","[4:1]a15"],
            ["[2:1]a13","[4:1]a14"], ["[2:1]a14,a18","[4:1]a20,a16"],
            ["[2:1]a16","[4:1]a18"], ["[2:1]a15,a17","[4:1]a17,a19"],
            ["[3:1]a1","[4:1]8R1a1"], ["[3:1]a2","[4:1]8R1a3"],["[3:1]a3","[4:1]8R1a5"],
            ["[3:1]a4,a8","[4:1]8R1a6,8R1a10"], ["[3:1]a5,a7","[4:1]8R1a7,8R1a9"],
            ["[3:1]6","[4:1]8R1a8"]
        ],
    },
    {   "ID": "[c]",
        "atomMap": [
            ["[4:1]a4","[0:1]a1"], ["[4:1]a1","[0:1]a2"],["[4:1]a3","[0:1]a10"],["[4:1]a2","[0:1]a3"],
            ["[4:1]a5","[0:1]a4"],["[4:1]a6","[0:1]a5"],["[4:1]a12","[0:1]a12"],["[4:1]a11","[0:1]a11"],
            ["[4:1]a13","[0:1]a13"],["[4:1]7R1a2","[0:1]6R1a2"],["[4:1]7R1a3","[0:1]6R1a3"],
            ["[4:1]7R1a6","[0:1]6R1a6"],["[4:1]7R1a5","[0:1]6R1a5"],["[4:1]7R1a9","[0:1]6R1a10"],
            ["[4:1]7R1a7","[0:1]6R1a7"],["[4:1]7R1a8R1a4","[0:1]6R1a8R1a4"],
            ["[4:1]7R1a8R1a3","[0:1]6R1a8R1a3"],["[4:1]7R1a8R1a2","[0:1]6R1a8R1a2"],
            ["[4:1]7R1a10","[0:1]6R1a9"],["[4:1]8R1a1","[0:1]7R1a2"],["[4:1]8R1a3","[0:1]7R1a3"],
            ["[4:1]8R1a1","[0:1]7R1a2","[0:1]7R2a2"],["[4:1]8R1a3","[0:1]7R1a3","[0:1]7R2a3"],
            ["[4:1]8R1a5","[0:1]7R1a5","[0:1]7R2a5"],
            ["[4:1]8R1a6,8R1a10","[0:1]7R1a6,7R1a10","[0:1]7R2a6,7R2a10"],
            ["[4:1]8R1a7,8R1a9","[0:1]7R1a7,7R1a9","[0:1]7R2a7,7R2a9"],
            ["[4:1]8R1a8","[0:1]7R1a8","[0:1]7R2a8"],
            ["[4:2]a2","[0:1]7R1a11"],["[4:2]a1","[0:1]7R1a12"],
            ["[4:1]a9","[0:1]a8"],["[4:1]a10","[0:1]a9"],["[4:1]a15","[0:1]a15"],["[4:1]a14","[0:1]a14"],
            ["[4:1]a16,a20","[0:1]a16,a20"],["[4:1]a17,a19","[0:1]a17,a19"],["[4:1]a18","[0:1]a18"]
        ],
    }
  ]
}
```

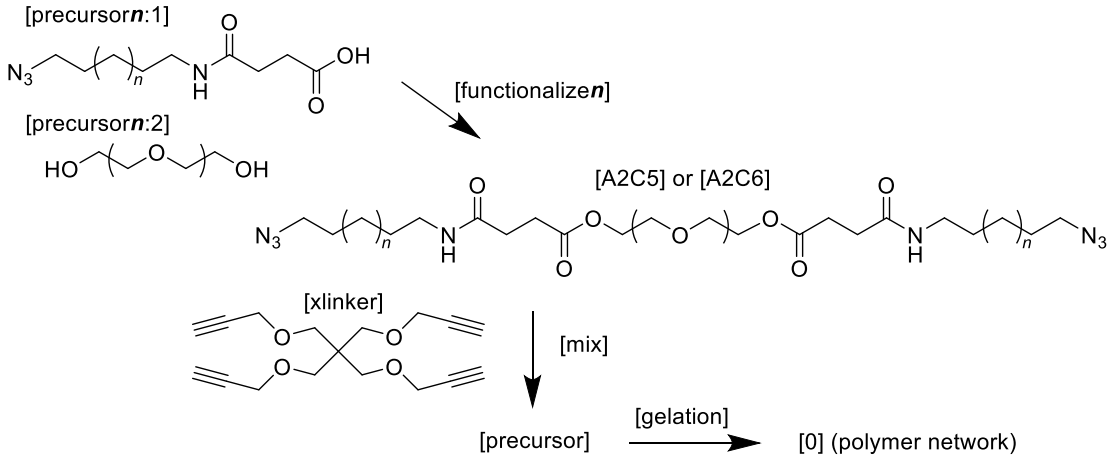
Polymer

Network disassembly spectroscopy of polymer network.

Source

Kawamoto, K., Zhong, M., Wang, R., Olsen, B. D., & Johnson, J. A. (2015). Loops versus branch functionality in model click hydrogels. *Macromolecules*, *48*(24), 8980-8988.

Schematics



Description & Notes

In this example, network disassembly spectroscopy data is captured using the proposed schema. Note that the fraction of unlabeled A_2 strands (A2C5) and labeled strands (A2C6) are denoted as variables *x* and *1-x* respectively within the schema to make the presentation more aligned with that of the source article. Similarly, the concentrations for the five distinct junction states (1111,1112,1122,1222 and 2222) are also denoted with variables *c1111*, *c1112*, *c1122*, *c1222*, *c2222*. In real instances, these variables will be replaced by the corresponding values found in the experiments.

PolyDAT


```
{
   "preamble": {
      "polymer": "{[][<][C:11]CCCCNC(=O)CCC(=O)O[C:12]{[$][$]COC[$][$]}[C:13]O
                     C(=O)CCC(=O)NCCCC[C:14][<],
                     [<][C:21]CCCCCNC(=O)CCC(=O)O[C:22]{[$][$]COC[$][$]}[C:23]O
                     C(=O)CCC(=O)NCCCCC[C:24][<],
                     C(COCC1=CN([>])N=N1)(COCC1=CN([>])N=N1)(COCC1=CN([>])N=N1)COC
                     C1=CN([>])N=N1[]}",
      "pdVersion": 1, "mixfileVersion": 0.01, "docID": "conjugate-01",
      "logs": [
         {   "author_id": ["ORCID: https://orcid.org/0000-0002-7510-1194"],
             "date": "2020-09-06", "msg": "Created Entry" }      ],
      "srcs": [
         {   "citeID": "src1", "doi": "10.1021/acs.macromol.5b02243",
             "desc": "NDS." }
      ],
      "network": [ "[precursor1]>[functionalize1]>[A2C5]",
                   "[precursor2]>[functionalize2]>[A2C6]",
```

```
                "[A2C5].[A2C6].[xlink]>[mix]>[precursor]",
                "[precursor]>[gelation]>[0]" ]
},
"species": [
    {   "ID": "[precursor1]",
        "contents": [
            {   "ID": "[precursor1:1]",
                "bigsmiles": "[N-]=[N+]=N[C:1]CCCCNC(=O)CCC(=O)O"
            },
            {   "ID": "[precursor1:2]",
                "bigsmiles": "O[C:1]{[$][$]COC[$][$]}[C:2]O",
                "characterization": {
                    "Mn": [ { "value":4600, "unit":"UO_0000221", "src":"src1" } ],
                    "Mw/Mn": [ { "value":1.02, "unit":"UO_0000186", "src":"src1" } ]
                }
            }
        ]
    },
    {   "ID": "[precursor2]",
        "contents": [
            {   "ID": "[precursor2:1]",
                "bigsmiles": "[N-]=[N+]=N[C:1]CCCCCNC(=O)CCC(=O)O"
            },
            {   "ID": "[precursor2:2]",
                "bigsmiles": "O[C:1]{[$][$]COC[$][$]}[C:2]O",
                "characterization": {
                    "Mn": [ { "value":4600, "unit":"UO_0000221", "src":"src1" } ],
                    "Mw/Mn": [ { "value":1.02, "unit":"UO_0000186", "src":"src1" } ]
                }
            }
        ]
    },
    {   "ID": "[A2C5]",
        "contents": [
            {   "ID": "[A2C5:1]",
                "bigsmiles": "[N-]=[N+]=N[C:1]CCCCNC(=O)CCC(=O)O[C:2]{[$][$]COC[$][$]}
                             [C:3]OC(=O)CCC(=O)NCCCC[C:4]N=[N+]=[N-]"
            }
        ]
    },
    {   "ID": "[A2C6]",
        "contents": [
            {   "ID": "[A2C6:1]",
                "bigsmiles": "[N-]=[N+]=N[C:1]CCCCCNC(=O)CCC(=O)O[C:2]{[$][$]COC[$][$]}
                             [C:3]OC(=O)CCC(=O)NCCCCC[C:4]N=[N+]=[N-]"
            }
        ]
    },
    {   "ID": "[xlinker]",
        "contents": [
```

```
        {   "ID": "[xlinker:1]",
            "bigsmiles": "C(COCC#C)(COCC#C)(COCC#C)COCC#C",
        }
    ]
},
{   "ID": "[precursor]",
    "contents": [
        {   "ID": "[precursor:A2C5]",
            "bigsmiles": "[N-]=[N+]=N[C:1]CCCCNC(=O)CCC(=O)O[C:2]{[$][$]COC[$][$]}
                         [C:3]OC(=O)CCC(=O)NCCCC[C:4]N=[N+]=[N-]"
            "quantity": x, "unit": "UO_0010006"
        },
        {   "ID": "[precursor:A2C6]",
            "bigsmiles": "[N-]=[N+]=N[C:1]CCCCCNC(=O)CCC(=O)O[C:2]{[$][$]COC[$][$]}
                         [C:3]OC(=O)CCC(=O)NCCCCC[C:4]N=[N+]=[N-]"
            "quantity": 1-x, "unit": "UO_0010006"
        }
        {   "ID": "[precursor:xlinker]",
            "bigsmiles": "C(COCC#C)(COCC#C)(COCC#C)COCC#C",
            "quantity": 1, "unit": "UO_0010006"
        }
    ]
},
{   "ID": "[0]",
    "contents": [
        {   "ID": "[0:1]",
            "bigsmiles": "{[][<][C:11]CCCCNC(=O)CCC(=O)O[C:12]{[$][$]COC[$][$]}[C:13]O
                          C(=O)CCC(=O)NCCCC[C:14][<],
                          [<][C:21]CCCCCNC(=O)CCC(=O)O[C:22]{[$][$]COC[$][$]}[C:23]O
                          C(=O)CCC(=O)NCCCCC[C:24][<],
                          C(COCC1=CN([>])N=N1)(COCC1=CN([>])N=N1)(COCC1=CN([>])N=
                          N1)COCC1=CN([>])N=N1[]}"
            "characterization": {
                "ratios": [
                    {
                        "substructure" :  [
                            "C(COCC1=CN(N=N1)CCCCC[ND2])(COCC1=CN(N=N1)CCCCC
                             [ND2])(COCC1=CN(N=N1)CCCCC[ND2])COCC1=CN(N=N1)CCCCC[
                             ND2]",
                            "C(COCC1=CN(N=N1)CCCCCC[ND2])(COCC1=CN(N=N1)CCCCC
                             [ND2])(COCC1=CN(N=N1)CCCCC[ND2])COCC1=CN(N=N1)CCCCC[
                             ND2]",
                            "C(COCC1=CN(N=N1)CCCCCC[ND2])(COCC1=CN(N=N1)CCCCCC
                             [ND2])(COCC1=CN(N=N1)CCCCC[ND2])COCC1=CN(N=N1)CCCCC[
                             ND2]",
                            "C(COCC1=CN(N=N1)CCCCCC[ND2])(COCC1=CN(N=N1)CCCCCC
                             [ND2])(COCC1=CN(N=N1)CCCCCC[ND2])COCC1=CN(N=N1)CCCCC
                             [ND2]",
                            "C(COCC1=CN(N=N1)CCCCCC[ND2])(COCC1=CN(N=N1)CCCCCC
                             [ND2])(COCC1=CN(N=N1)CCCCCC[ND2])COCC1=CN(N=N1)CCCCC
                             C[ND2]"
```

```
                                ],
                                "ratio" : [ c1111, c1112, c1122, c1222, c2222 ],
                                "unit": "UO_0000013",
                                "src": "src1",
                                "method": "NDS"
                            }
                        ]
                    }
                }
            ]
        }
    ],
    "transformation": [
        {   "ID": "[functionalize1]",
            "atomMap": [
                [ "[precursor1:1]1", "[A2C5:1]1,4" ],
                [ "[precursor1:2]1,2", "[A2C5:1]2,3" ]
            ]
        },
        {   "ID": "[functionalize2]",
            "atomMap": [
                [ "[precursor2:1]1", "[A2C6:1]1,4" ],
                [ "[precursor2:2]1,2", "[A2C6:1]2,3" ]
            ]
        },
        {   "ID": "[mix]",
            "atomMap": [
                [ "[A2C5:1]1,4", "[precursor:A2C5]1,4" ],
                [ "[A2C5:1]2,3", "[precursor:A2C5]2,3" ],
                [ "[A2C6:1]1,4", "[precursor:A2C6]1,4" ],
                [ "[A2C6:1]2,3", "[precursor:A2C6]2,3" ]
            ]
        },
        {   "ID": "[gelation]",
            "atomMap": [
                [ "[precursor:A2C5]1,4", "[0:1]11,14" ],
                [ "[precursor:A2C5]2,3", "[0:1]12,13" ],
                [ "[precursor:A2C6]1,4", "[0:1]21,24" ],
                [ "[precursor:A2C6]2,3", "[0:1]22,23" ]
            ]
        }
    ]
}
```

Polymer

Melt Blending of silica nanoparticle and homopolymer.

Source

Moncada, E., Quijada, R., & Retuert, J. (2007). Nanoparticles prepared by the sol–gel method and their use in the formation of nanocomposites with polypropylene. *Nanotechnology*, 18(33), 335606.

Description & Notes

Synthesis of polypropylene nanocomposite.

PolyDAT

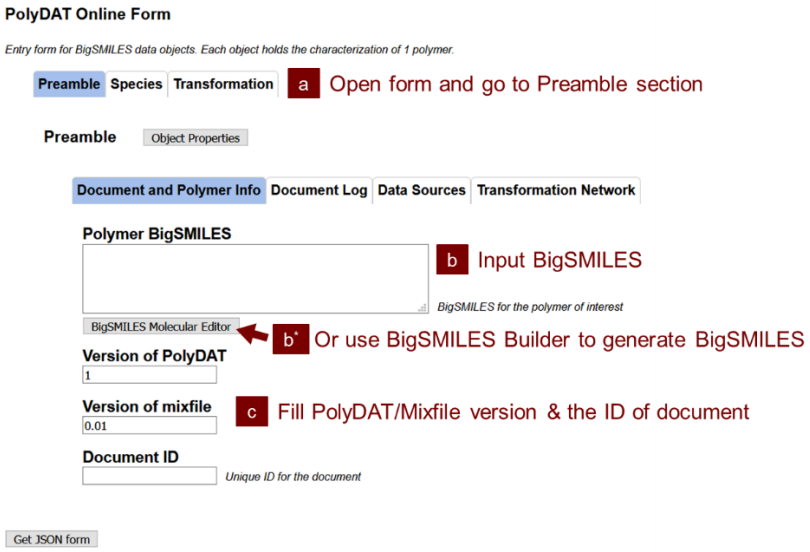

```
{
   "preamble": {
      "name": "Polypropylene-silica nanocomposite",
      "polymer": "{[][$]CC(C)[$][]}",
      "pdVersion": 1, "mixfileVersion": 0.01, "docID": "PP-SNP-1",
      "logs": [
         {   "author": ["ORCID:https://orcid.org/0000-0002-7272-7140"],
             "date": "2020-07-01", "msg": "Created Entry" }      ],
      "srcs": [
         {   "citeID": "src1", "doi": "10.1088/0957-4484/18/33/335606",
             "desc": "Nanoparticles prepared by the sol–gel method and their use in the formation of
               nanocomposites with polypropylene" }        ],
      "network": [ "[solution1].[solution2]>[reaction1]>[SNP]",
                   "[PP-g-MA].[BHT].[Irganox].[SNP]>[mixing]>[materbatch]",
                   "[masterbatch].[PP].[BHT].[Irganox]>[blending]>[0]"  ]
   },
   "species": [
      {   "ID": "[0]",
         "contents": [
            { "ID": "[0:1]", "bigsmiles": "{[][$]CC(C)[$][]}"    },
            { "ID": "[0:2]", "bigsmiles": "{[][$]CC(C)[$],[$]CC(C)(C1C(=O)OC(=O)C1)[$][]}" },
            { "ID": "[0:3]", "empirical_formula": "SiO2", "name": "silica nanoparticle"   },
         ]
      },
      {   "ID": "[solution1]",
         "contents": [
            { "ID": "[solution1:H2O]", "bigsmiles": "O", "quantity": 54, "units": "UO_0000098" },
            { "ID": "[solution1:NH3]", "quantity": 2.2, "units": "UO_0000098"
              "contents" [     {   "ID":"[solution1:NH3:NH3]", "bigsmiles": "N",
                                   "quantity":28, "units":"UO_0000163" },
                               {   "ID":"[solution1:NH3:H2O]", "bigsmiles": "O"}    ]
            },
            { "ID": "[solution1:ethanol]", "quantity": 23.5, "units": "UO_0000098"
              "contents" [     {   "ID":"[solution1:ethanol:ethanol]", "bigsmiles": "CCO",
                                   "quantity":95.6, "units":"UO_0000165" },
                               {   "ID":"[solution1:ethanol:H2O]", "bigsmiles": "O"}  ]
            },
         ]
      },
```

```
        {   "ID": "[solution2]",
            "contents": [
                {   "ID": "[solution2:TEOS]", "bigsmiles": "CCO[Si](OCC)(OCC)OCC",
                    "quantity": 55, "units": "UO_0000098"    },
                {   "ID": "[solution2:ethanol]", "quantity": 23, "units": "UO_0000098"
                    "contents" [  {   "ID":"[solution1:ethanol:ethanol]", "bigsmiles": "CCO",
                                      "quantity":95.6, "units":"UO_0000165" },
                                  {   "ID":"[solution1:ethanol:H2O]", "bigsmiles": "O"}
                    ]
                },
            ]
        },
        {   "ID": "[SNP]", "description": "spherical silica nanoparticle", "contents": [
            { "ID":"[SNP:silica]", "empirical_formula": "SiO2" }   ]          }
        {   "ID": "[BHT]", "contents": [
            {"ID":"[BHT:1]", "bigsmiles": "CC(C)(C)c1cc(C)cc(c1O)C(C)(C)C "  }   ]       }
        {   "ID": "[Irganox]", "name": "Irganox", "description": "antioxidant"  }
        {   "ID": "[masterbatch]", "description": "mixture of compatibilizer and nanoparticles"  }
        {   "ID": "[PP-g-MA]", "contents": [
            {"ID":"[PP-g-MA:1]", "bigsmiles": "{[][$]CC(C)[$],[$]CC(C)(C1C(=O)OC(=O)C1)[$][]}" }
        ]   }
        {   "ID": "[PP]",
            "contents": [
                {   "ID": "[PP:1]"
                    "bigsmiles": "{[][$]CC(C)[$][]}",
                    "characterization": {
                        "Mw": [ { "value": 340, "unit": "UO_0000222" } ],
                        "Mw/Mn": [ { "value": 3.9, "unit": "UO_0000186" } ]
                    }
                }
            ]
        },
    ],
    "transformation": [
        {   "ID": "[reaction1]" },
        {   "ID": "[mixing]",
            "recipe" [{ "ingredient": "[PP-g-MA]", "quantity": 27, "units": "UO_0000021" },
                       { "ingredient": "[BHT]", "quantity": 0.02, "units": "UO_0000021" },
                       { "ingredient": "[Irganox]", "quantity": 0.02, "units": "UO_0000021" },
                       { "ingredient": "[SNP]", "quantity": 28.2, "units": "UO_0000098" }    ]
        }
        {   "ID": "[blending]",
            "recipe" [{ "ingredient": "[masterbatch]", "quantity": 1.48, "units": "UO_0000021" },
                       { "ingredient": "[BHT]", "quantity": 0.02, "units": "UO_0000021" },
                       { "ingredient": "[Irganox]", "quantity": 0.02, "units": "UO_0000021" },
                       { "ingredient": "[PP]", "quantity": 35.5, "units": "UO_0000021" }       ]
        }
    ]
}
```

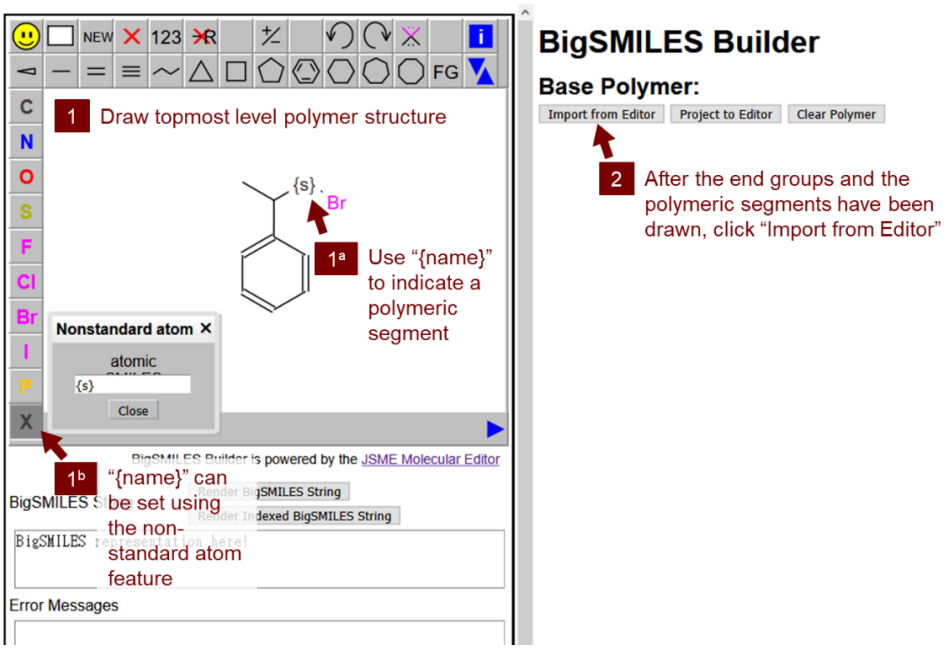
S7. Tutorial for PolyDAT Web Form and BigSMILES Builder

The PolyDAT web form can be accessed at https://olsenlabmit.github.io/PolyDAT-form/index.html. To showcase the utility of the program, a simple run through over how the form should be filled out is provided. A video version of this tutorial can be found at https://olsenlabmit.github.io/BigSMILES/.

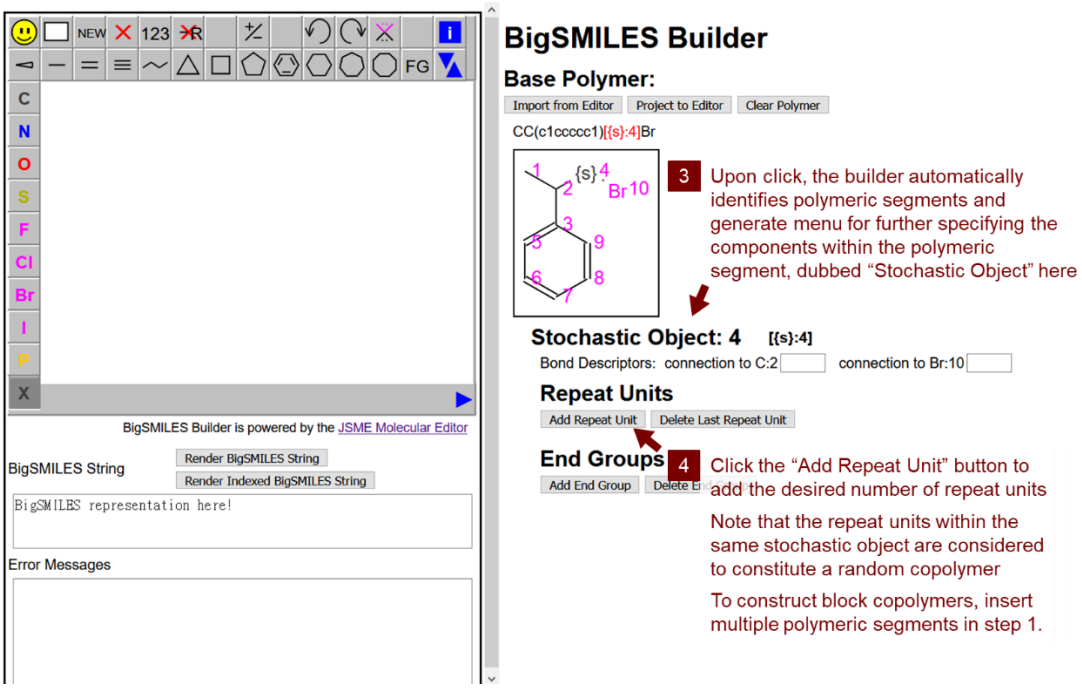
To begin, open the form and select the *Preamble* tab. Upon opening the tab, metadata entries such as the polymer BigSMILES, PolyDAT version, Mixfile version and the document ID are visible:



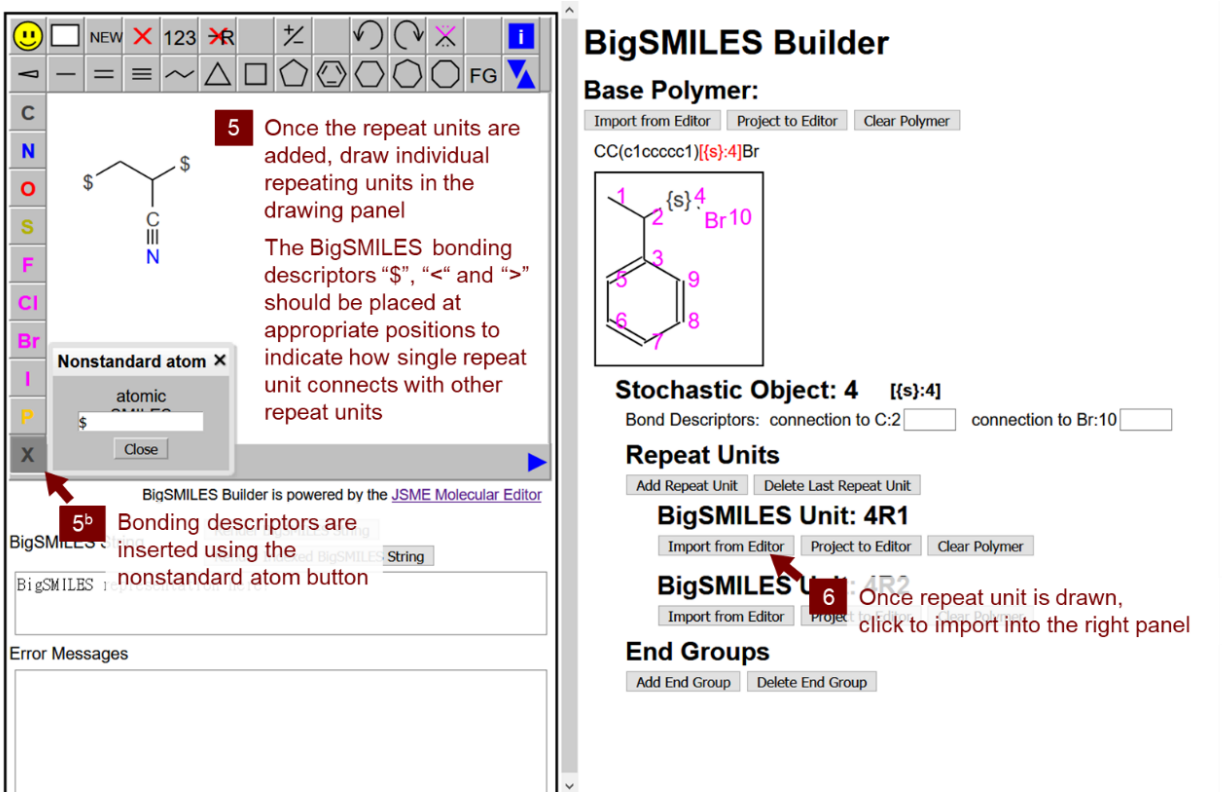
For users familiar with BigSMILES, the BigSMILES section could be filled out by hand. However, to make the form more accessible, a graphical interface is also available upon clicking the "BigSMILES Molecular Editor" button below the text area. Once clicked, a popup window will appear. To use the editor, the structure of the polymer should be specified in a hierarchical manner. Consider a linear copolymer of styrene and acrylonitrile as example. Firstly, the topmost level polymer structure should be specified. For the copolymer considered, as illustrated in the following figure, a stochastic object, represented by the nonstandard atom "{Z}", that serves as a placeholder for a polymeric segment should be drawn first, followed by the specification of the end groups at the two ends of the polymeric segment. Note that within the editor, any nonstandard atom that is bracketed by a pair of curly brackets will be picked up by the program as stochastic objects.



Once the top-level structure of the polymer is sketched, the structure can be imported into the right panel by clicking the import button. Upon import, the program will automatically identify any embedded stochastic object. If a stochastic object is picked up by the program, a new panel will appear to let the user further specify the component of the polymeric segment:

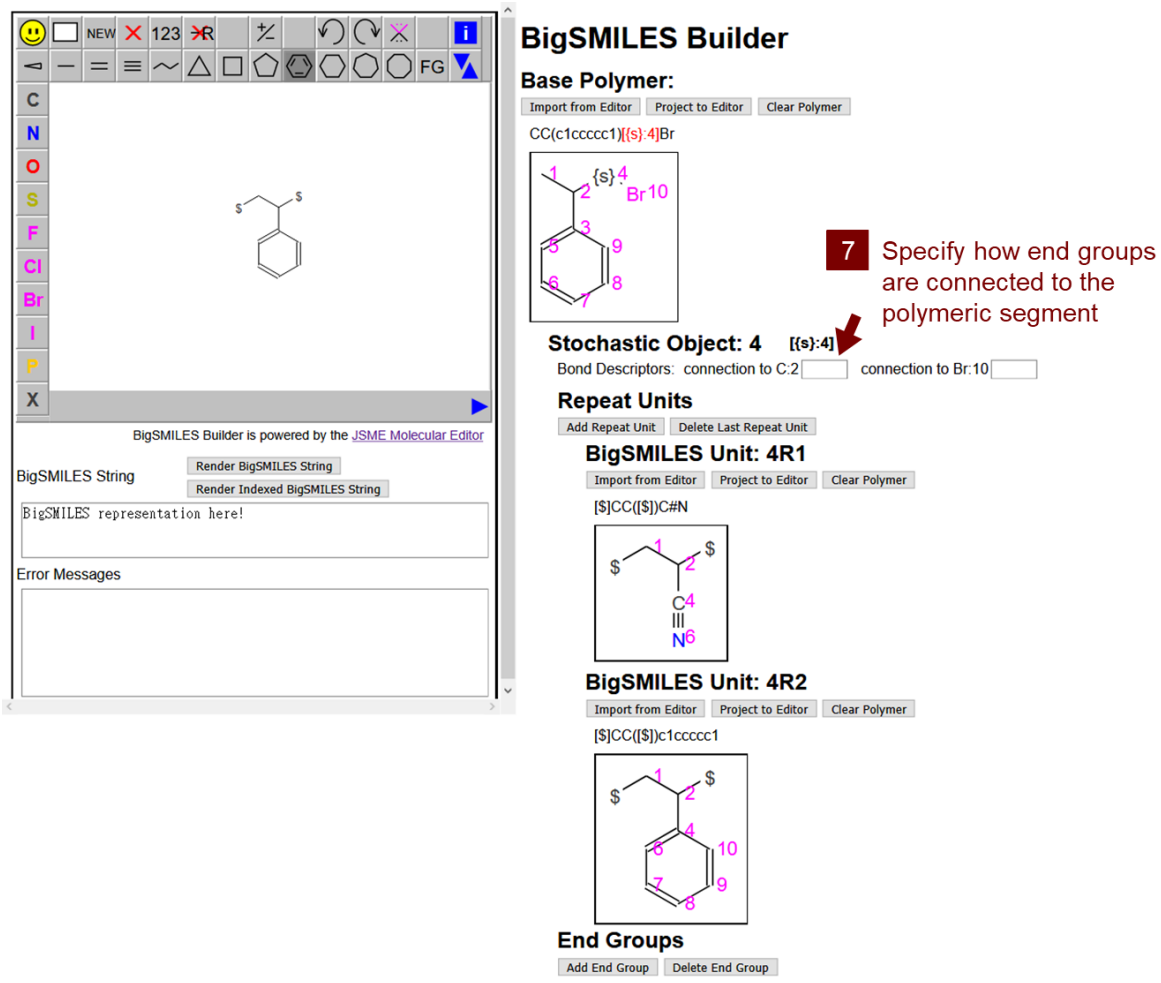


To specify the internal structure of the polymeric unit, click on the "Add Repeat Unit" button to insert new repeat units. Then, draw the repeat units in the left panel and import each one into the right panel by clicking the import button on the corresponding object. Note that each repeat unit should contain at least two BigSMILES bonding descriptor sites that indicates where the repeat unit interconnects with other repeat units. The bonding descriptors can be either "$", "<" or ">". The "$" symbol indicates a connection site that could be joined with any other "$" site to form a bond. This is typically seen in vinyl polymers. In contrast, the "<" sites can only be joined by the opposite ">" sites and vice versa. Therefore, ">" and "<" are commonly found on repeat units that undergo condensation reactions. If multiple orthogonal connections need to be specified, a trailing integer can be added to the bonding descriptors, e.g. "$1", to distinguish between chemically orthogonal groups. Like the stochastic objects, the bonding descriptors can be inserted by using the nonstandard atom button in the drawing panel.

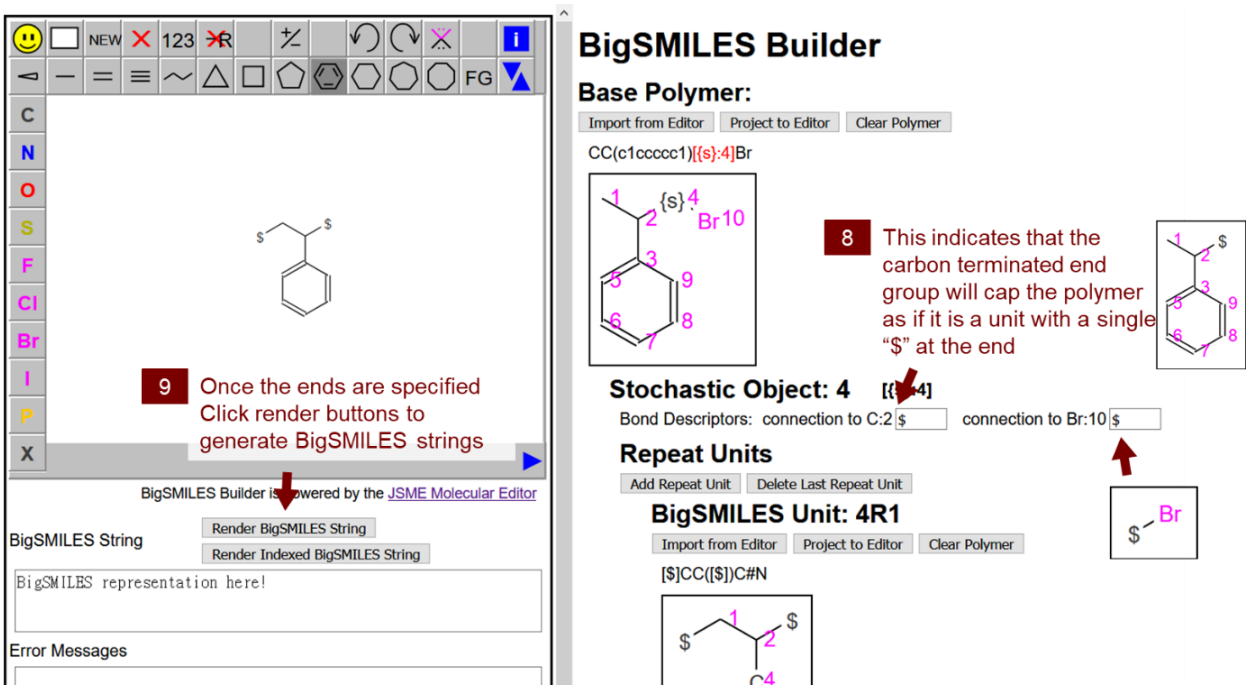


Once the repeat unit is fully sketched, it can be imported into the right panel using the "Import from Editor" button. Note that macromonomers are supported, and repeat units can contain additional stochastic objects.

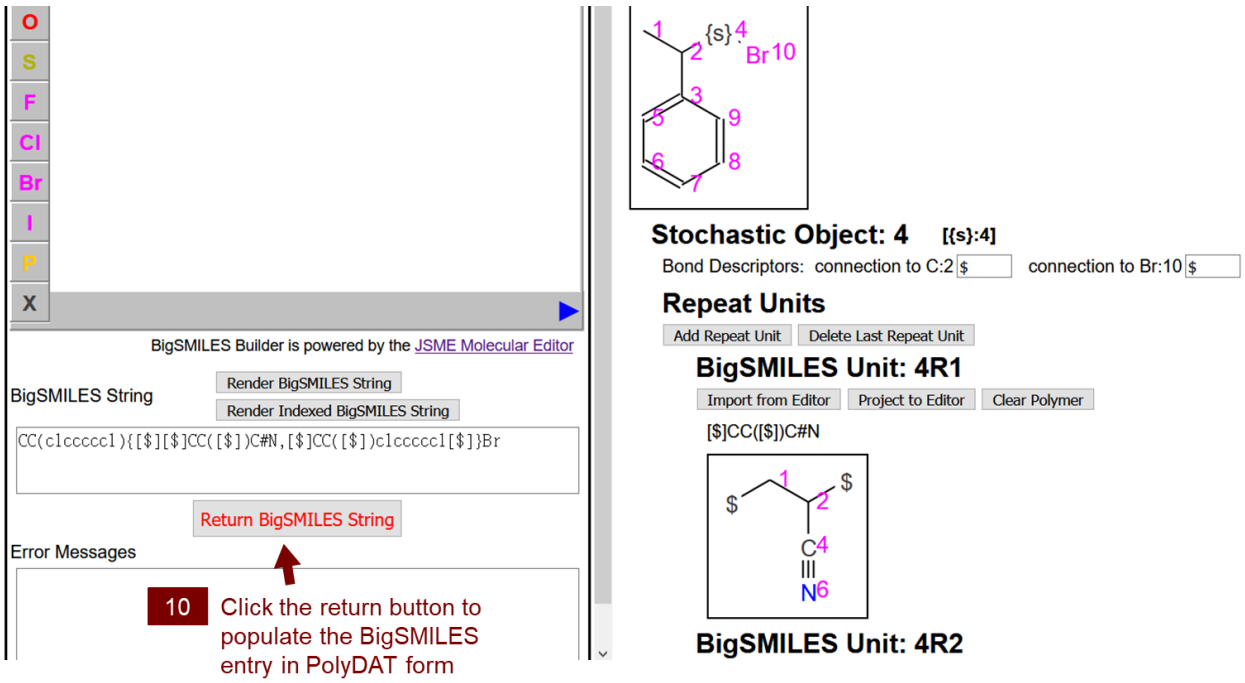
In this case, the program will similarly pick up the stochastic objects and generate another layer of panels of repeat units within the macromolecular repeat unit. Once all the repeat units have been specified, how the polymeric unit connects to the exterior should be specified. This is specified by filling bonding descriptor patterns at the input boxes indicated in the following figure.



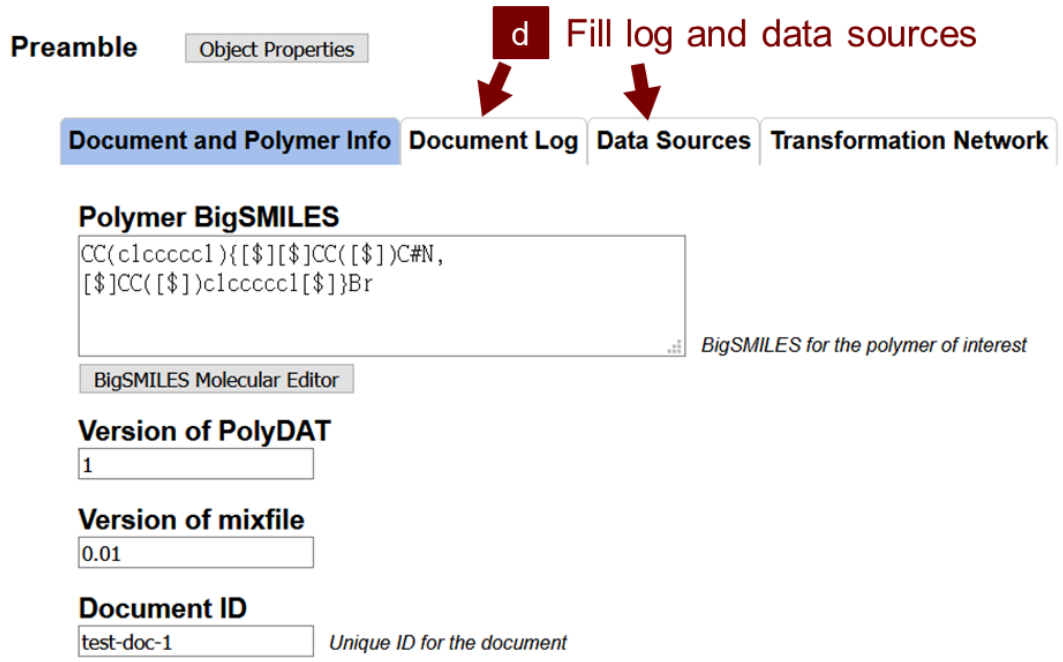
Note that the specified terminal bonding descriptors represent the type of bonding descriptor that would be attached to the terminal end groups if they were to be viewed as a "repeat unit" with a single bonding site. In the example illustrated in the following figure, the first "$" indicates that the carbon end of the polymeric segment is terminated as if the benzylic end group act as a unit with a single "$" descriptor attached; meanwhile, the second "$" indicates that the other end of the polymer is terminated as if the bromine end group is connected to a single "$" descriptor:



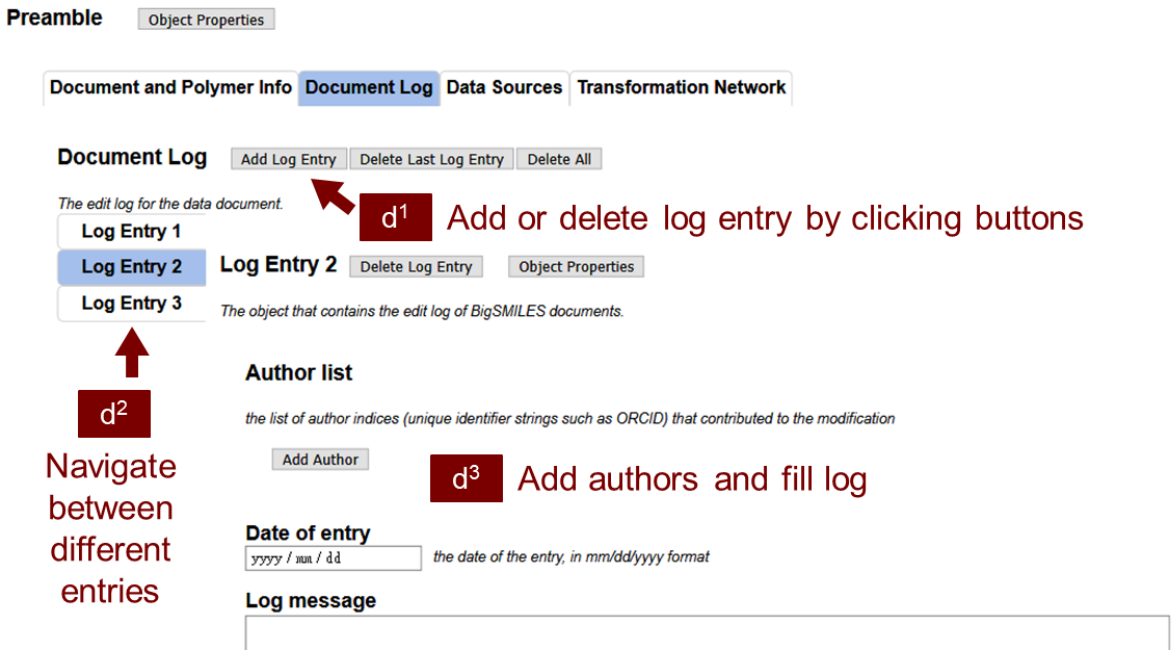
Once the entire polymer is specified, the BigSMILES string is rendered by clicking the rendering buttons in the left panel. Finally, after the string is rendered, the user will be prompted to return to the original PolyDAT web from.



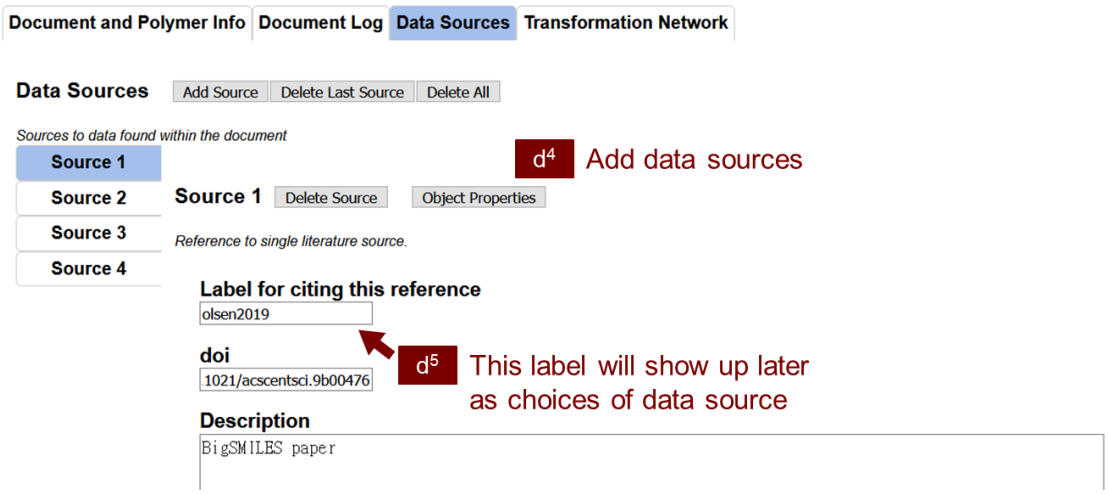
The BigSMILES entry will be automatically populated upon returning to the web form. Next, the log and the data source sections should be filled out.



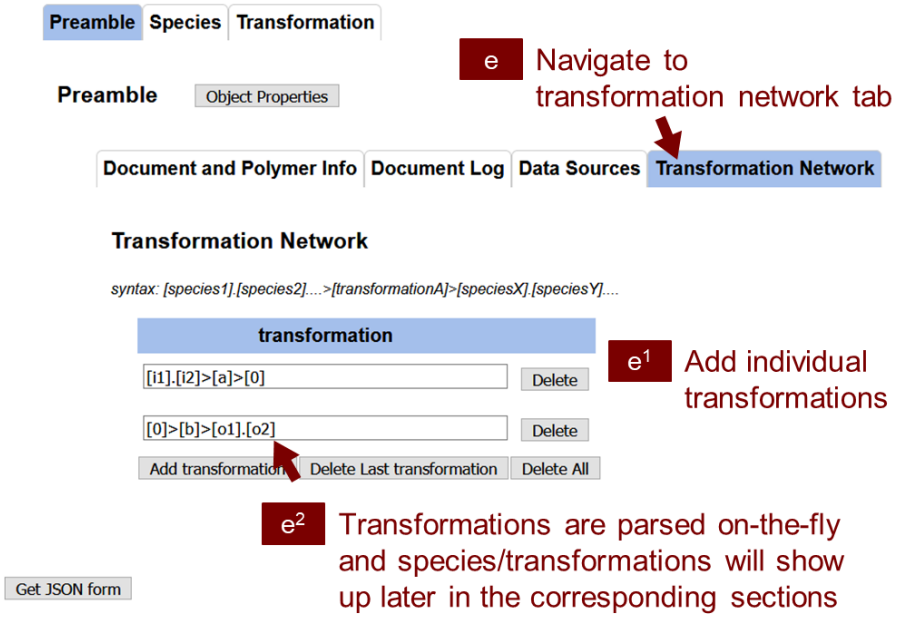
Log entries can be manipulated by invoking the corresponding buttons:



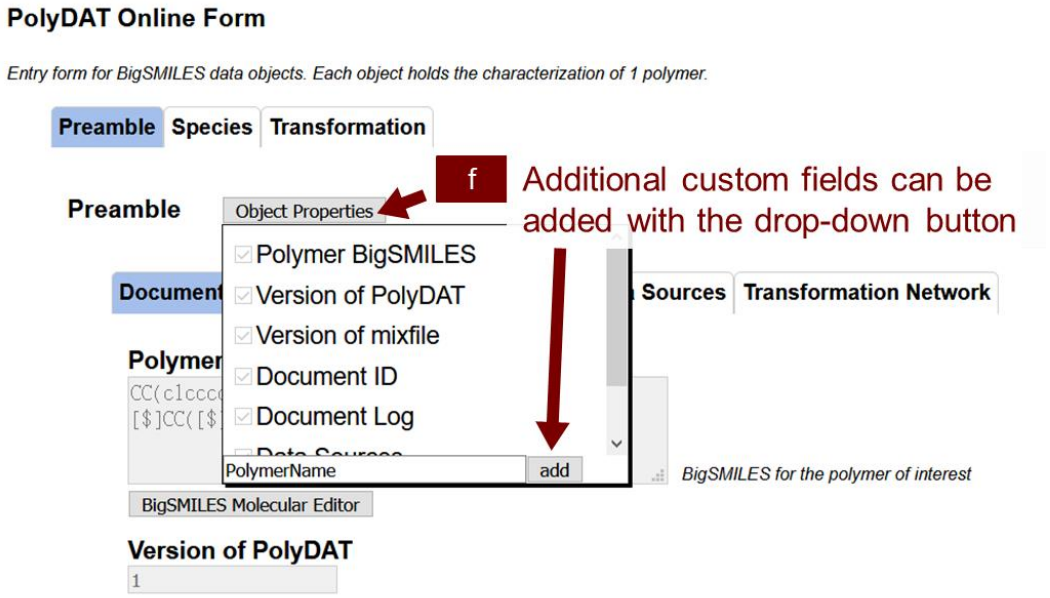
Likewise, the data source tab can be filled out in a similar manner. It should be noted that unique identifiers should always be used as the labels for each data source, as the labels entered in this section are parsed and used as the options in the other sections of the form.



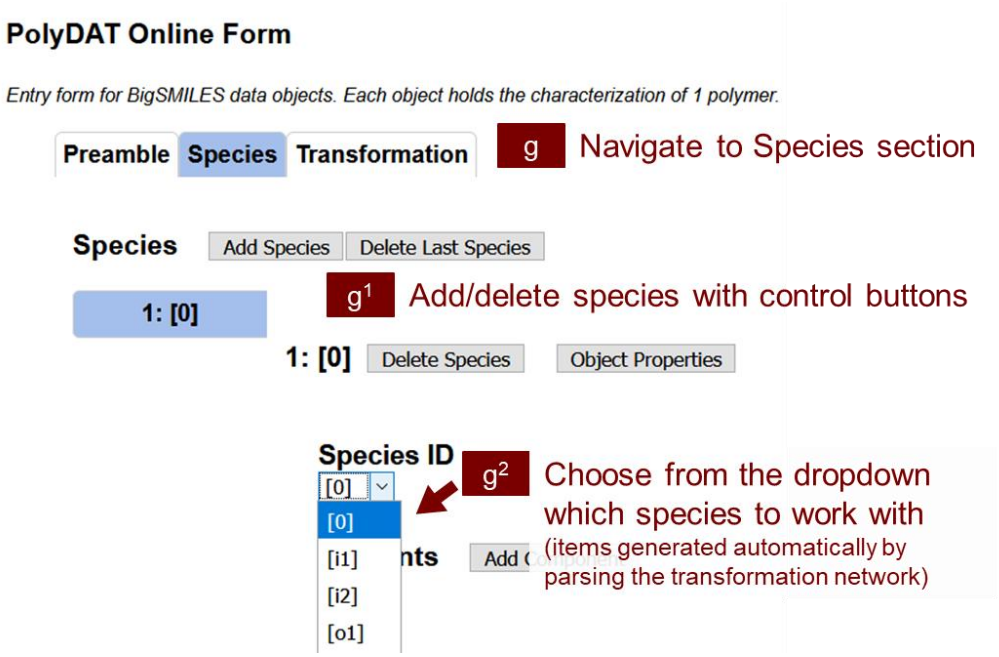
The final reaction network section of the preamble can be filled out straightforwardly. Note that the transformations specified in this section are parsed and used as dropdown options for the *Species* and *Transformation* sections. Therefore, this section should always be filled out prior to the other two sections to avoid incorrect behavior.



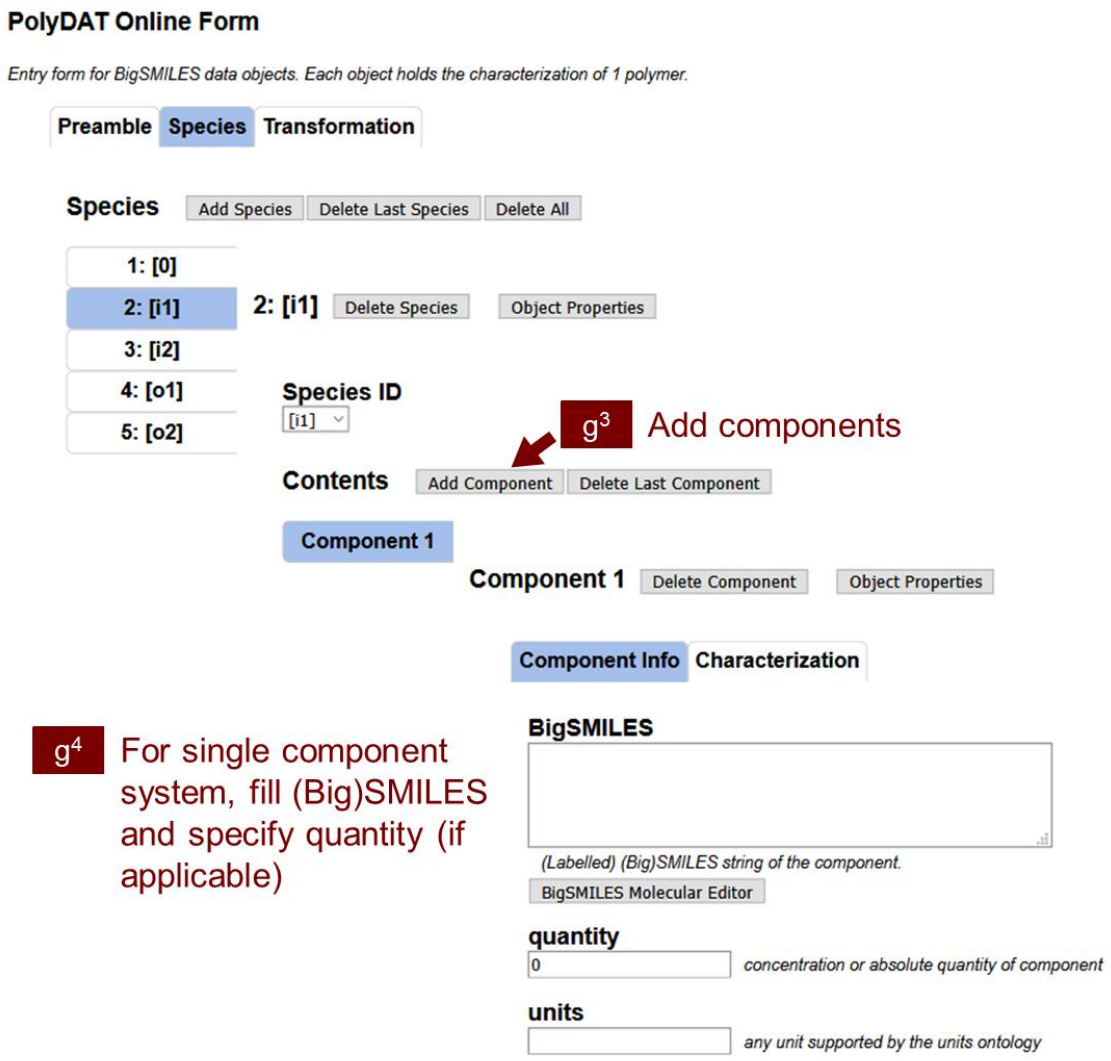
Before moving on to the other sections, it should be noted that additional custom fields can be added to the preamble section by invoking the drop-down menu indicated in the following figure.



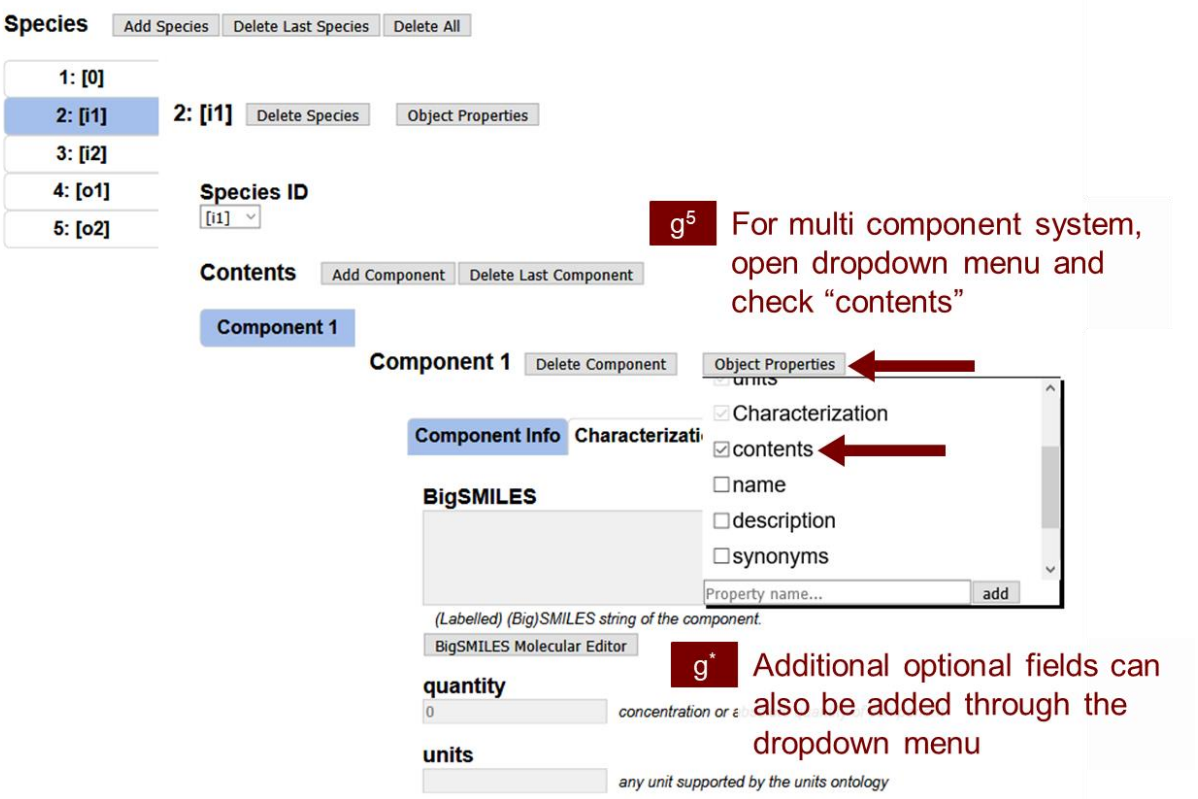
To specify the relevant species, navigate to the *Species* tab. Within the tab, individual species can be added or deleted by the provided controls:



The contents of each species can be specified by adding components to the species. For components that constitute of a single substance, the BigSMILES for the component should be filled along with the quantity of the component.

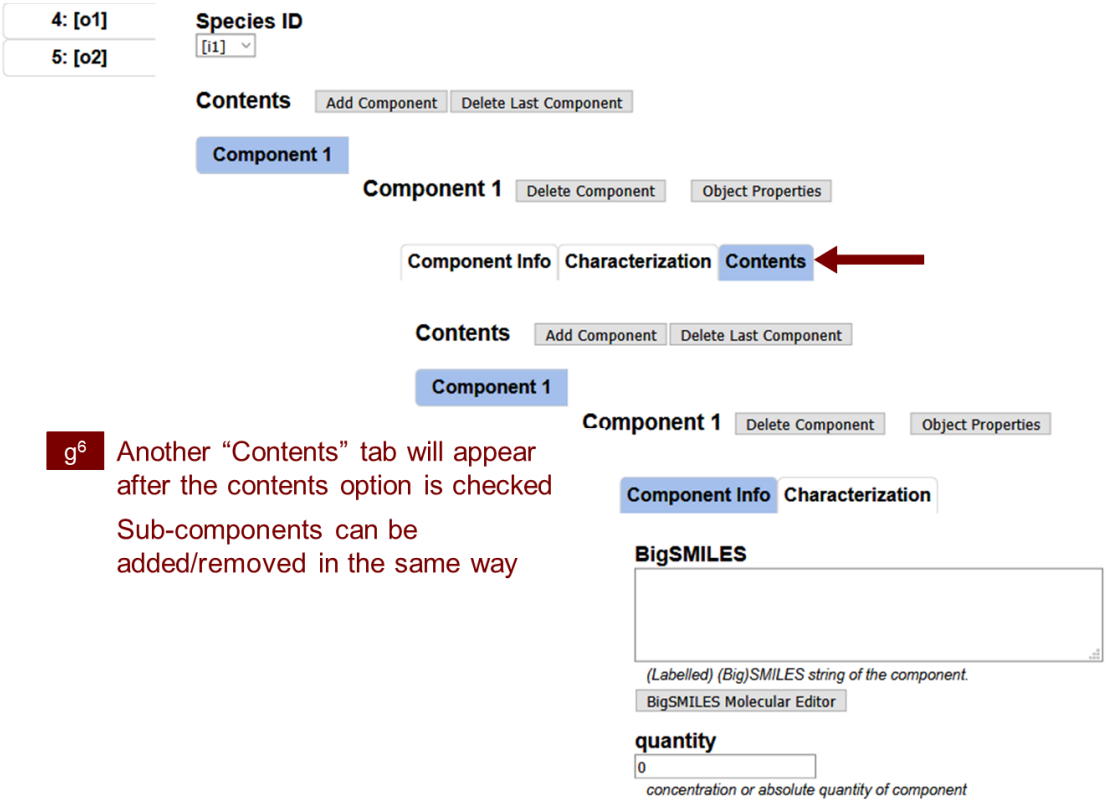


For components that are mixtures of multiple sub-components, the composition should be specified by detailing the composition for each of the sub-components under another level of *contents*.



g5 For multi component system, open dropdown menu and check “contents”

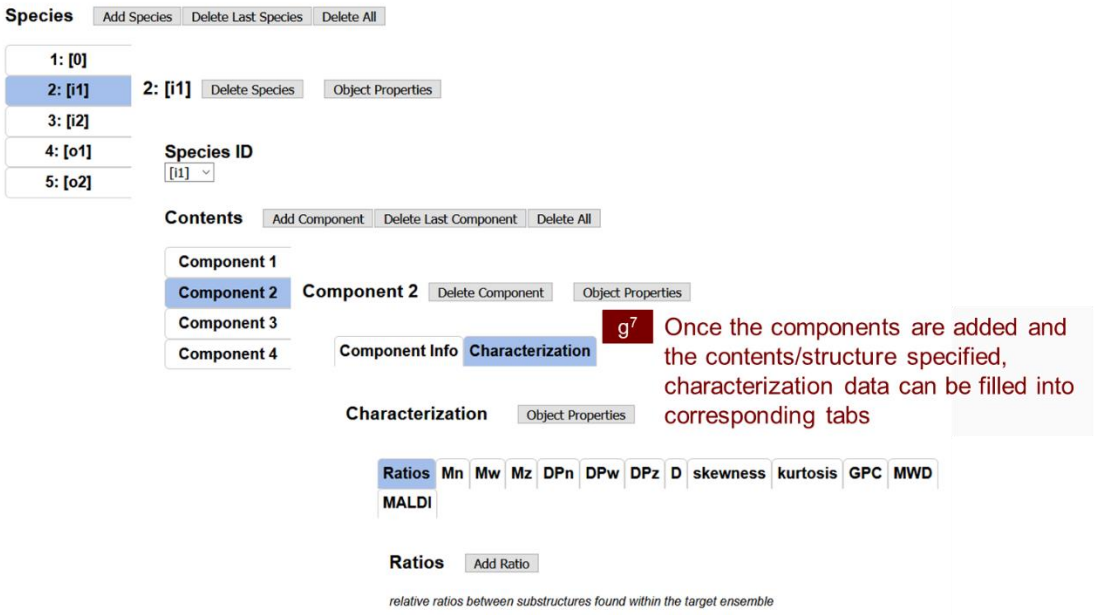
g* Additional optional fields can also be added through the dropdown menu



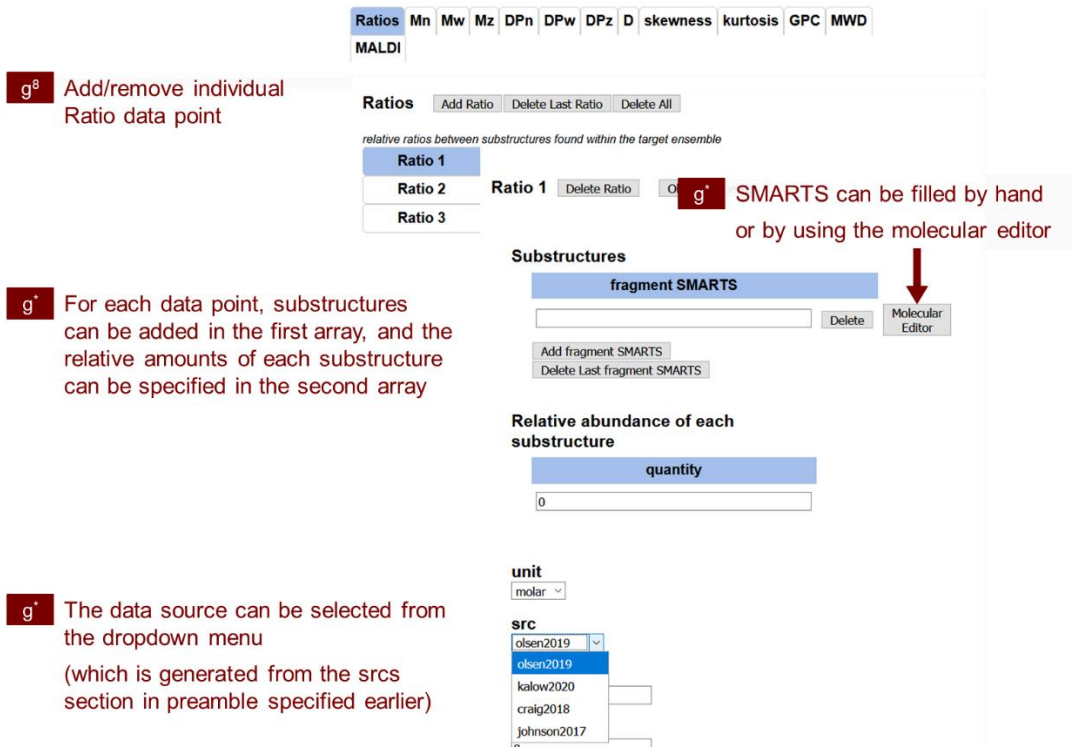
g6 Another “Contents” tab will appear after the contents option is checked

Sub-components can be added/removed in the same way

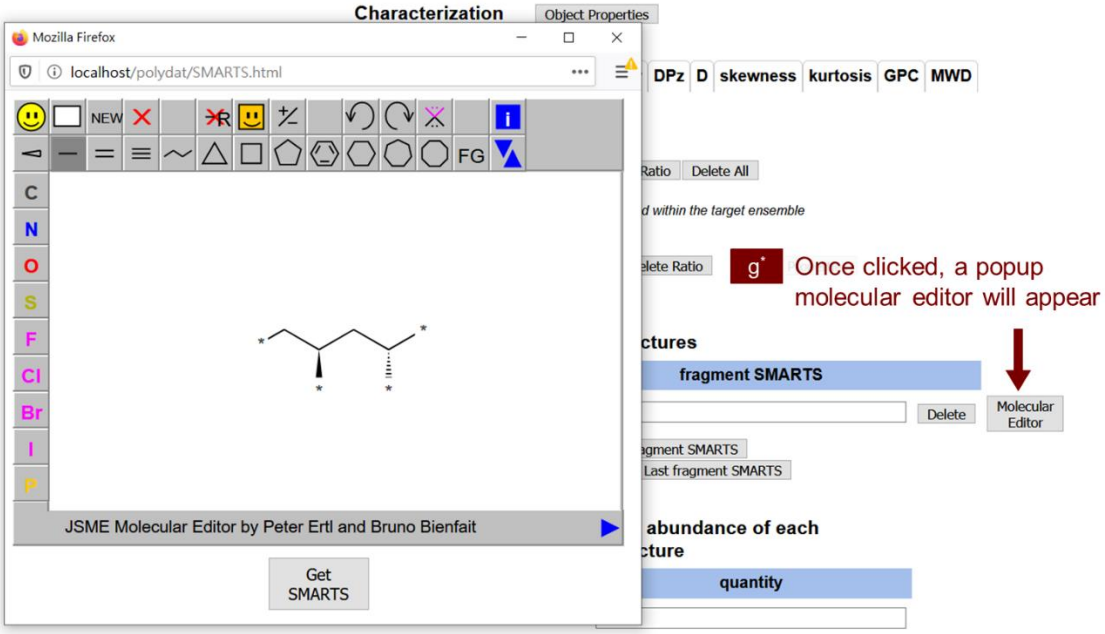
In addition to the specification of the chemical composition, characterization data can also be attached to a specific component under the *Characterization* tab.



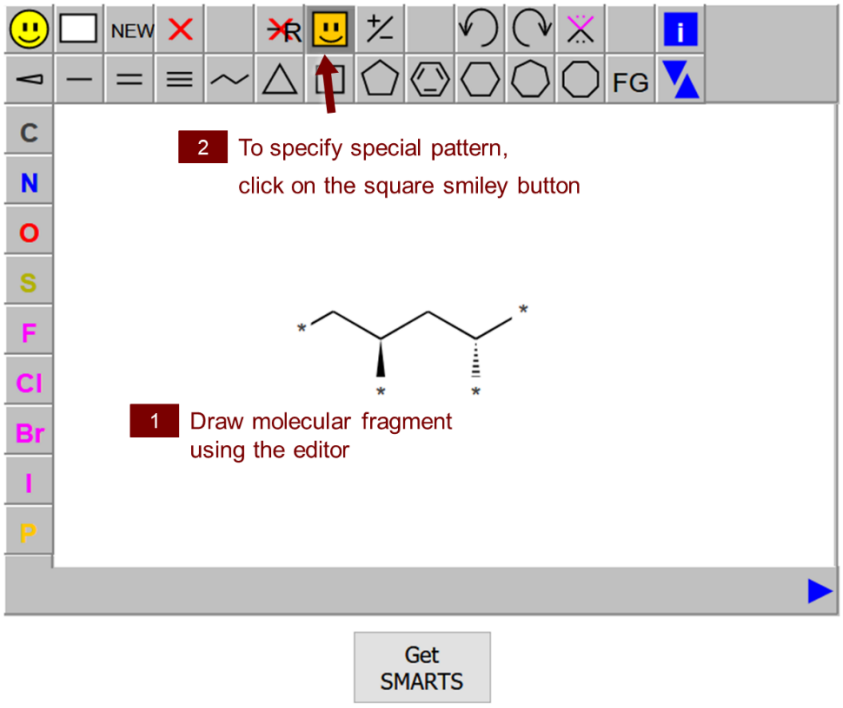
Individual measurements and data points should be added under the corresponding tab.



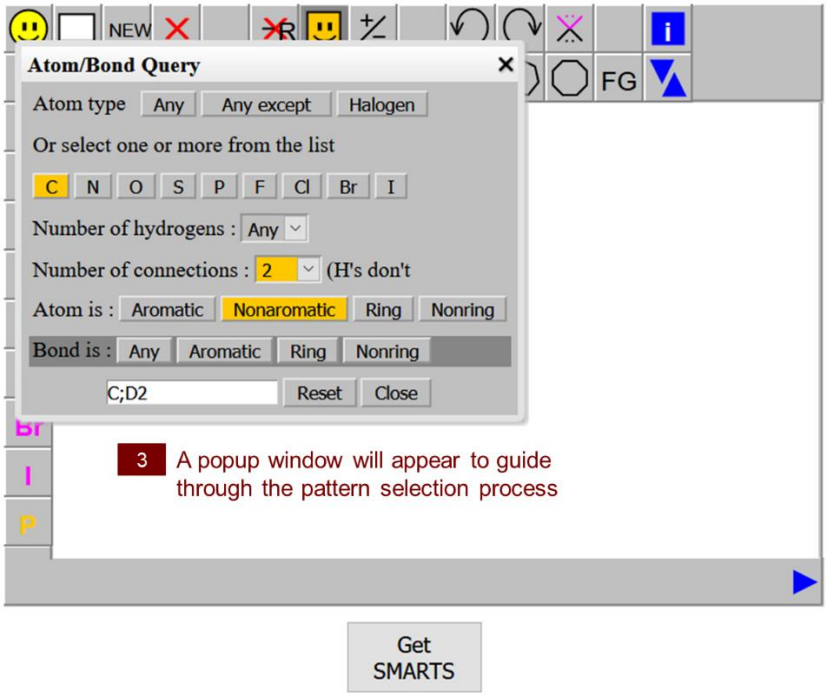
Within the *Ratios* tab, a molecular editor that helps with the construction of SMARTS patterns can be accessed by clicking the "Molecular Editor" button.



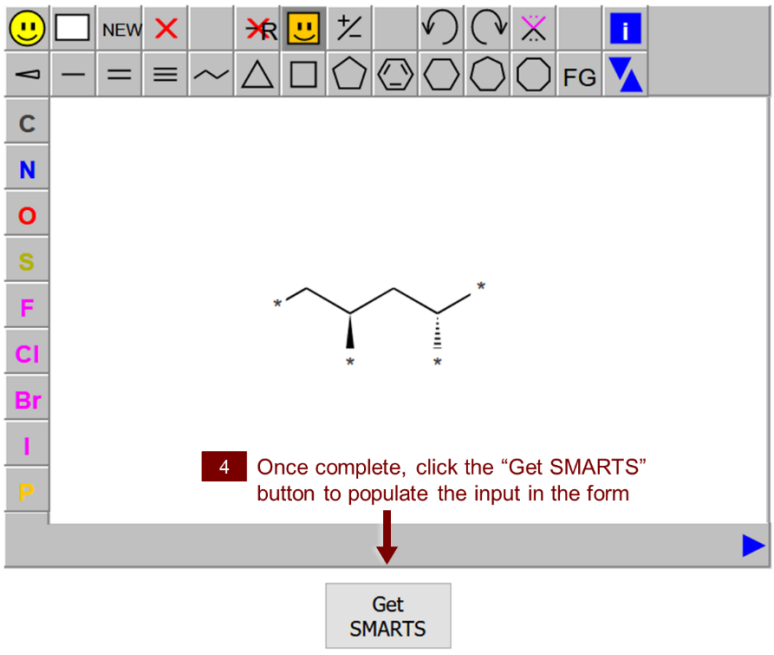
Within the editor, usual atoms/bonds can be placed using normal editor inputs.



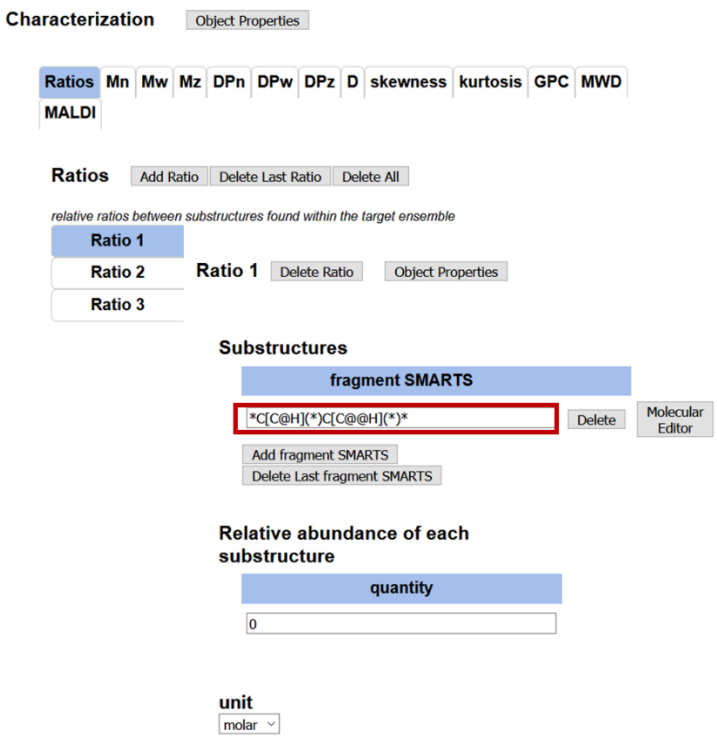
Meanwhile, the smiley button at the top panel can be clicked to bring out a popup menu providing options for inputting SMARTS patterns for atoms and bond:



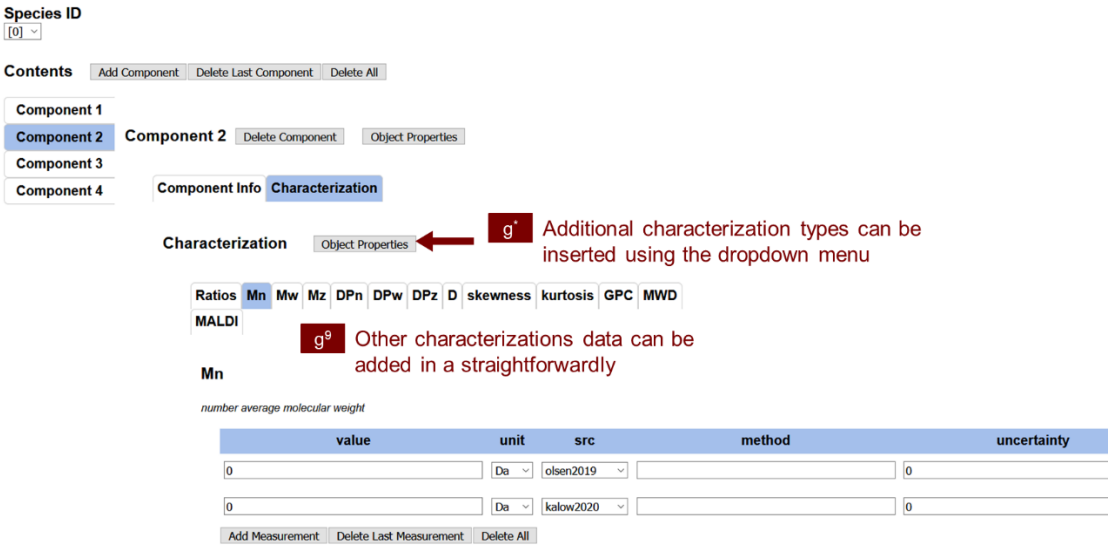
Finally, once the desired SMARTS pattern has been drawn, the bottom button can be clicked to close the editor and return to the PolyDAT form.



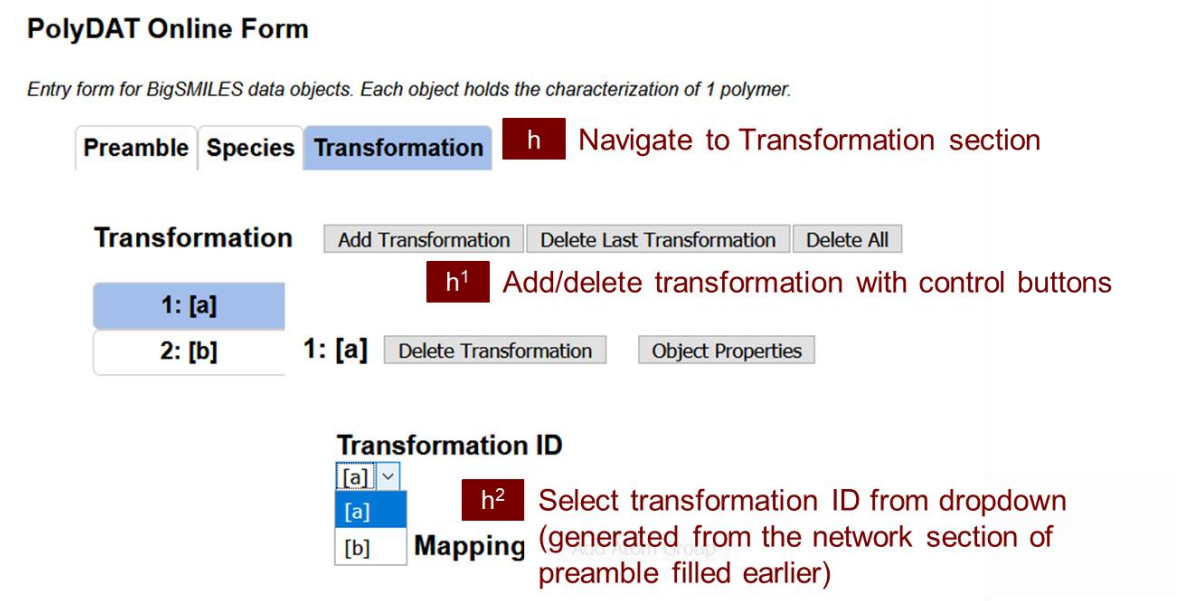
Similar to the BigSMILES builder, the generated SMARTS will automatically populate the corresponding entry upon returning to the web form.

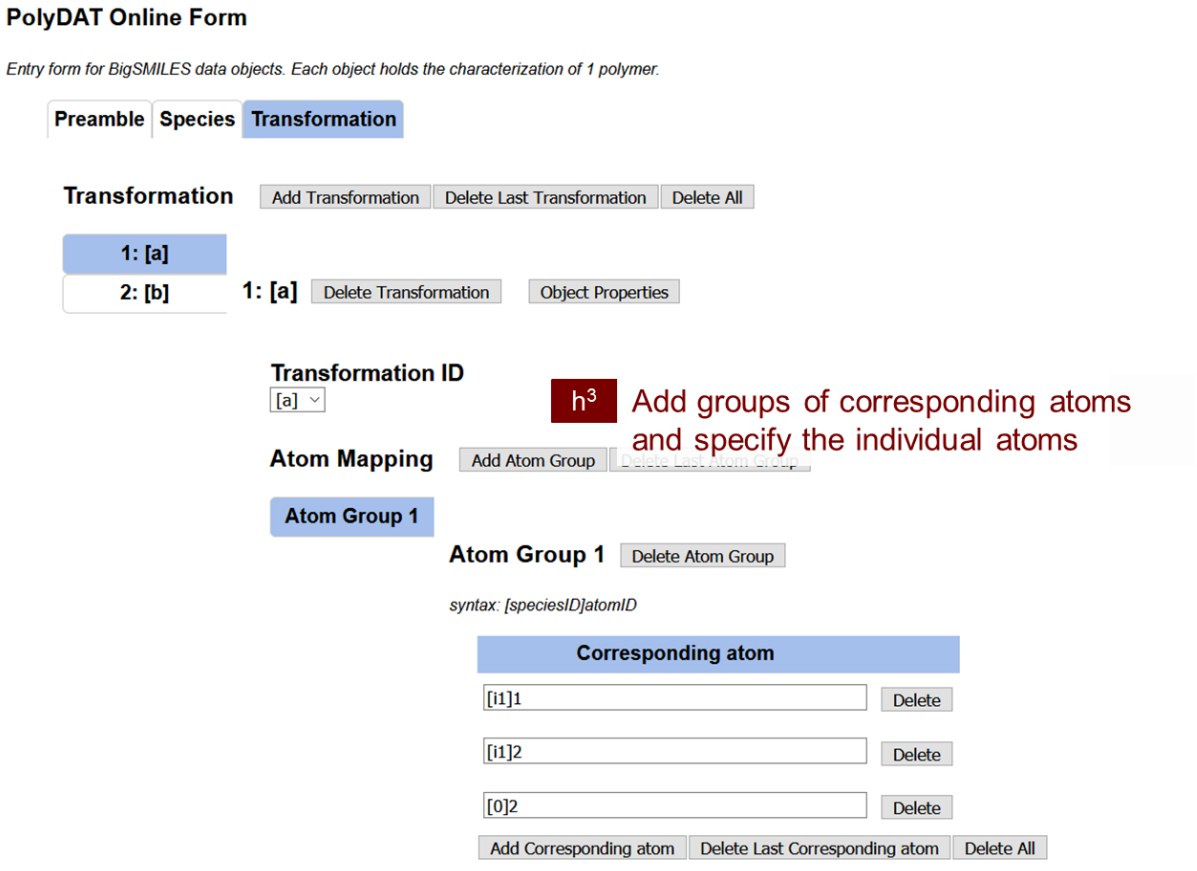


Like the *Ratios* section, other sections can also be filled in a similar manner

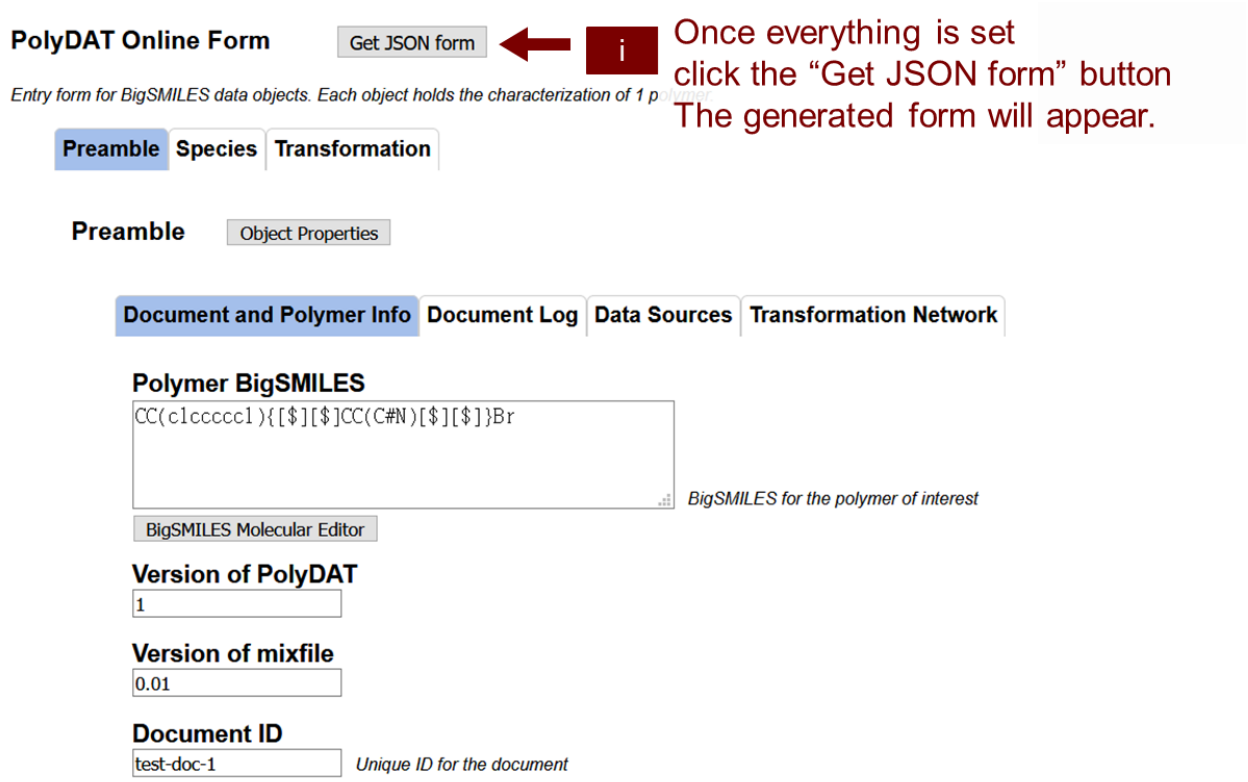


After the *Species* section is filled, the *Transformation* section can be accessed and filled in a similar manner.





Finally, once the entire form has been completed, the "Get JSON form" button at the top of the page can be invoked to generate the encoded PolyDAT object in JSON format.



S8. Text Copies for PolyDAT Examples in Figures

Figure 2

Figure 2a


```
{
    "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" : [ ... ]
}
```


Figure 2b


```
{
    "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 4


```
{   "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" : [ ... ]
}
```

Figure 5

```
"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 6

```
"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]#[N:12])[$],
                [$][C:13][C:14]([c:15]1[c:16][c:17][c:18][c:19][c:20]1)[$][$]}[Br:21]",
"characterization" : {
    "ratios": [
        {   "substructure" : [ "[C:9][C:10]", "[C:13][C:14]" ],
            "ratio" : [ 0.80, 0.20 ], "unit"  : "UO_0000013"   },
        {   "substructure" : [ "[#6]", "[#7]" ],
            "ratio" : [ 5, 1 ], "unit"  : "UO_0000013"  },
        {   "substructure" : [ "[CD2][CD3][CD2]", "[CD2][CD3][CD3]" ],
            "ratio" : [ 0.90, 0.10], "unit"  : "UO_0000013"    },
        {   "substructure": [ "C[C@H]([#6])CC[C@H]([#6])[*]",
                              "C[C@H]([#6])CC[C@@H]([#6])[*]" ],
            "ratio" : [ 0.55, 0.45], "unit"  : "UO_0000013"    }
    ]
    "Mn": [{ "value": 10, "unit": "UO_0000222", "uncertainty": 0.2,
             "method": { "methodName" : "osmometry", … } }],
    "Mw": [{ "value": 14, "unit": "UO_0000222", "uncertainty": 2,
              "method": { "methodName" : "GPC", …} }],
    "Mw/Mn": [{ "value": 1.35, "unit": "UO_0000186", "method": {…}, "uncertainty": 0.1 }],
    "MWD": [ "y-value": [ 1,3,10, …], "x-value": [0.1,0.2,0.3,…], "x-unit": "UO_0000222" ]
}
```

Figure 7

```
{
    "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 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 }
        ]
    }
]
```

References

1. Clark, A. M.; McEwen, L. R.; Gedeck, P.; Bunin, B. A., Capturing mixture composition: an open machine-readable format for representing mixed substances. *Journal of cheminformatics* **2019,** *11* (1), 33.

2. Gkoutos, G. V.; Schofield, P. N.; Hoehndorf, R., The Units Ontology: a tool for integrating units of measurement in science. *Database* **2012,** *2012*.