# Program for generating new polymer structures

The program has the following limitations:

1) Allowed chemical elements in the composition of the polymer: C, H, N, O, F, Si, S, Cl, Br

2) To indicate a repeating unit, two star atoms are used. Framework structures, graft and block copolymers, spatial polymers (when several asterisks must be specified to indicate a repeating unit) are not processed.

3) Each star must be connected by a single bond to a single atom

4) Polymers with isotopes are not processed - all isotope marks are nulled before processing

The program consists of three blocks: the formation of a database of fragments of the main chain and substituents, the block of prediction of properties and the choice of properties for prediction, the generation of polymer structures

## Prediction of polymer properties.

Bicerano regression models [r6] are used for prediction. The Askadsky model [r7] is used to predict water permeability. Below is a list of properties that can be predicted

Table 1. List of properties predicted by PolyProp and their abbreviations

|  |  |
| --- | --- |
| Abbreviation | Description |
| 0CHI | Zero-order chi |
| 0CHIV | Zero order chiV |
| 1CHI | First order chi |
| 1CHIV | First order chiV |
| NA | Number of atoms, including H |
| NONH | Number of non-hydrogen atoms |
| MWPRU | Molweight PRU Molecular weight of PRU |
| VDW | Van-der-Waals volume |
| V | Molar volume |
| MW | Molweight of polymer (average?) |
| D | Density Density |
| CL | Head capacity, liquid |
| CS | Heat capacity solid |
| COH1 | Cohesion energy, Feudor |
| COH2 | Cohesion energy, Van Crevelen |
| DELTA1 | Delta solubility, Feudor |
| DELTA2 | Delta solubility, Van Crevelen |
| N | Refractive index |
| RLL | Molar refraction |
| EPSILON | Dielectric constant |
| PLL | Molar polarizability |
| MU | Effective dipole moment |
| TG | Glass temperature for molweight pMW |
| TGINF | Glass temperature at infinite molweight |
| RAO | Molar Rao function |
| HARTMAN | Molar Hartman function |
| MB | Bulk modulus |
| ME | Young modulus Young modulus |
| MS | Shear modulus Shear modulus |
| ENTMW | Entaglement Molweight |
| POISSON | Poisson ratio |
| SIGMAFINF | Brittle fracture stress at infinite molweight |
| SIGMAF | Brittle fracture stress at molweight pMW |
| SIGMAY | Yield stress Yield stress |
| PRULENGTHUSED | Length PRU |
| STIFFNESS | Molar stiffness |
| MOLVISCOSITYFUNC | Molar viscosity function |
| EACTIVATIONFLOW | Activation energy of viscous flow |
| ZEROSHEARVISCOSITY | Zero shear viscosity |
| HEATCONDUCTIVITY | Heat conductivity |
| O2PERMEABILITY | Permeability O2 |
| N2PERMEABILITY | Permeability of N2 |
| CO2PERMEABILITY | Permeability of CO2 |
| H2OPERMEABILITY | Permeability of H2O |
| DECOMPOSITION | Decomposition function |
| DECOMPOSITIONTEMPERATURE | Decomposition temperature |
| DISSIPATIONLOW | Low-frequency dissipation energy |
| DISSIPATIONHIGH | High-frequency dissipation |
| SOLUBLE | List of solvents, in which a polymer is soluble |
| UNSOLUBLE | List of solvents, in which a polymer is unsoluble |

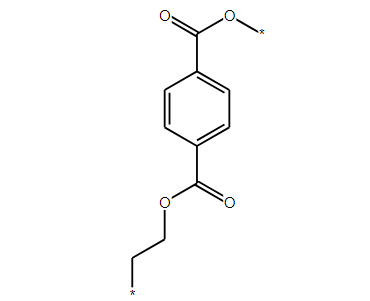
To predict properties, a string is formed consisting of abbreviated designations of properties (the first column of Table 1) and separated by spaces. The string is supplied as a command line parameter. If there is no property line, then all properties are predicted. If the prediction of a property requires the value of another property that is not included in the list of predictable properties, then that property is added to the list. Adding recursively, that is, a chain of dependent properties can be added to the list.

If the property string is empty or not included in the list of command parameters, then all properties are predicted.

For properties: V(Molar volume), D(Density), CL(Heat capacity of polymer in liquid phase), CS(Heat capacity of polymer in solid phase), N(Refractive index), MB(Volume modulus), MS(Shear modulus), ME(Young's Modulus), POISSON(Poisson's Ratio), SIGMAFINF(Brittle Fracture Stress at Infinite Molar Mass), SIGMAY(Yield Strength), HEATCONDUCTIVITY (Thermal conductivity), H2OPERMEABILITY (H2O permeability), ZEROSHEARVISCOSITY (Viscosity at zero shear) calculations of the temperature dependence of their values are provided.

For properties SIGMAF (Brittle Fracture Stress) and ZEROSHEARVISCOSITY (Viscosity at Zero Shear), molecular weight dependency calculations are possible.

The properties prediction results are displayed in XML files for further processing in the MultiComp software package and in an HTML file containing information for the user (Fig. 1)



|  |  |  |  |
| --- | --- | --- | --- |
| Property name | Abbreviation | Value | Units of measure |
| Connectivity index 0X | 0χ | 9.966255 |  |
| Connectivity index 0Xv | 0χv | 7.356608 |  |
| Connectivity index 1X | 1χ | 6.770857 |  |
| Connectivity index 1Xv | 1χv | 4.215214 |  |
| Repeat unit number of atoms | NAtoms | 22 |  |
| Repeat unit number of non-hydrogen atoms | NNonH | 14 |  |
| Molweight PRU | mw | 192.171120 | g/mole |
| Van-der-Waals Volume | Vw | 95.043400 | cm3/mole |
| Molar Volume | V | 145.906844 | cm3/mole |
| Polymer Molweight | mw | 100000.000000 | g/mole |
| Density | d | 1.317081 | g/cm3 |
| HeatCapacity(liquid) | C(l) | 311.476898 | J/mole/K |
| HeatCapacity(solid) | C(s) | 228.868423 | J/mole/K |
| Cohesion Energy Fedors | ECoh1 | 74086.000000 | J/mole |
| Cohesion Energy Van Krevelen | ECoh2 | 57031.144728 | J/mole |
| Solubility parameter Feudor | δ | 22.533582 | (J/cc)0.5 |
| Solubility parameter VanKrevelen | δ | 19.770525 | (J/cc)0.5 |
| Refractive Index | nD | 1.555804 |  |
| Molar Refractivity | RLL | 46.886865 | cm3/mole |
| Dielectric constant | ε | 3.282741 |  |
| Molar Polarizability | PLL | 63.048247 | cm3/mole |
| Effective dipole Moment | μ | 0.885737 | Debye |
| Glass temperature at molweight 100000.000000 | Tg | 369.530916 | K |
| Infinite molweight Tg | Tginf | 368.160911 | K |
| Bulk modulus | B | 4411.275179 | MPa |
| Young modulus | E | 2594.893883 | MPa |
| Shear modulus | G | 925.452316 | MPa |
| Entanglement Molweight | me | 5732.790343 | g/mole |
| Brittle fracture stress at infinite molweight | σf(∞) | 112.761375 | Mpa |
| Brittle fracture stress at molweight 100000.000000 | σf(∞) | 90.782506 | Mpa |
| Yield stress | σy | 72.657029 | Mpa |
| PRU length (used) | l | 7.189514 | A |
| Molar Rao function | UR | 8568.091706 | cm10/3/(sec1/3mole) |
| Molar Hartman function | UH | 6596.460723 | cm10/3/(sec1/3mole) |
| Poisson ratio | nu | 0.401960 |  |
| Molar stiffness function | K | 80.936657 | g0.25cm1.5/mole0.75 |
| Molar viscosity function | Hη | 7001.243871 | g\*J1/3\*mole-4/3 |
| Activation energy viscous flow | Eη | 48.357212 | kJ/mol |
| Thermal conductivity at 298K | λ(298) | 0.209668 | J/(K\*m\*sec) |
| Permeability of oxygen at 298K | PO2 | 0.039764 | Barrers |
| Permeability of nitrogen at 298K | PN2 | 0.007916 | Barrers |
| Permeability of carbon dioxide at 298K | PCO2 | 0.137064 | Barrers |
| Permeability of water vapours at 298K | PH2O | 588.952416 | Barrers |
| Molar Thermal Decomposition function | Yd,1/2 | 140.650619 | K\*kg/mol |
| Temperature of half decomposition | Td,1/2 | 731.903000 | K |
| Soluble in solvents |  | Morpholine, Pyridine, Cellosolve |  |
| Unsoluble in solvents |  | Ethylene glycol, Glycerol, Water |  |

Figure 1. HTML page showing results of property prediction for PET.

Below, a graphical interface for selecting properties for predictions will be described.

## Generation of new polymer structures.

Databases of substituents and fragments of the main chain are used to generate new polymer structures.

Generation begins with the selection of fragments of the main chain and substituents. Fragments must be in the database - new ones are not allowed. In the future, it is planned to add new fragments to the bases of the main chain and substituents. The graphical interface for selecting fragments will be described below.

When generating new structures, the main chain of the polymer is first generated and then substituents are added to it.

## Polymer backbone generation

To do this, the number of Nfrag fragments used to generate the main chain is set. Two algorithms for generating the main chain of polymers have been implemented - enumeration of all possible options and random selection of fragments (Monte Carlo). In the search algorithm for the main chain, all possible combinations of used fragments are built by the number Nfrag. In the Monte Carlo algorithm, the combination of fragments used is chosen randomly.

One of the main chain generation options is a variable number of Nfrag fragments. With this option enabled, NFrag is treated as the maximum number of fragments in the main chain. That is, the number of fragments varies from 1 to Nfrag. In the enumeration algorithm, all possible combinations of the number of fragments of the main chain are used, in the Monte Carlo algorithm, the number of fragments in the chain is chosen randomly from 1 to Nfrag.

Another ambiguity that arises during the generation of the main chain is the non-equivalence of the "head" and "tail" of the fragment.

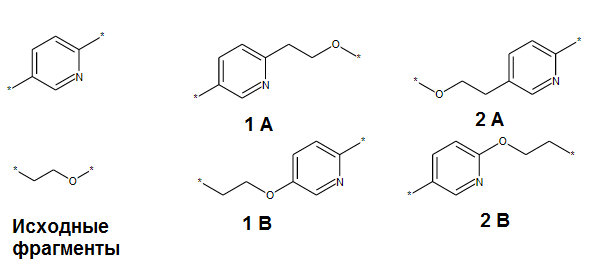


Figure 2. Initial fragments and two possible chains of polymers 1 and 2. Pairs of structures (1A and 1B), (2A and 2B) are equivalent, but 1 and 2 are different.

Figure 2 shows all possible chemical structures of polymers generated from two fragments. Actually generated two unique structures - 1 and 2. Structures A and B are equivalent.

When using the Monte Carlo algorithm, fragments of the main chain are selected randomly. Their connection points (head to tail, head to head, tail to tail) are also chosen randomly. Randomly selected substituents are attached at points of attachment. And then the next main chain is generated from random fragments, etc.

In the search algorithm, all possible substituents are searched at the attachment points, and this, as a rule, requires a large number of calculations. Therefore, it is important to avoid repetition of the chemical structures of the main chain in cases where this does not lead to the loss of generated polymers - namely, when using a brute-force algorithm to generate structures that do not contain a list of substituents for atoms in the main chain. If there are lists of substituents, then in this case it is necessary to take into account all possible combinations of structures in the main chain, including duplicates.

To understand this, let us consider the generation of polyhalophenylene structures. Figure 3 shows the numbering of atoms

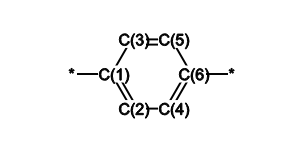


Figure 3. Numbering of atoms in para-phenylene.

The chlorine atom can attach to the 2,3,4,5 atoms and the fluorine atom to the 2,3 positions. It is also specified that the number of para-phenylene fragments in the repeating unit of the polymer is two. It is possible to generate two different PRUs (polymer repeated units) containing two fluorine atoms (Fig. 4)

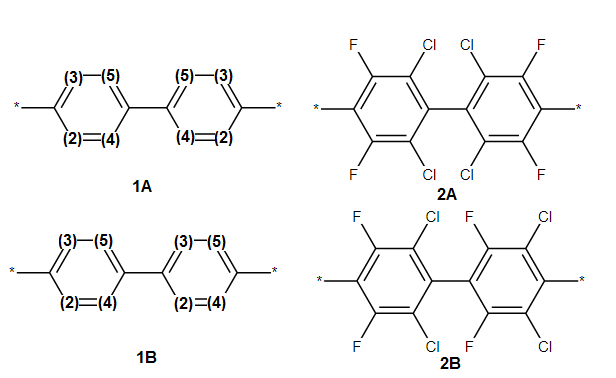


Figure 4. Two different repeating groups of polymers (2A, 2B). They were generated from formally identical polymer backbones (1A, 1B) but with different atom numbering.

That is, if there are lists of substituents for the atoms of the main chain, it is impossible to check the main chain for identity with previously used main chains of polymers. This check should not be carried out when using the Monte Carlo algorithm. But if polymers are generated by replacing all possible hydrogen atoms in the main chain, then such a check will avoid the generation of identical structures.

## Adding substitutes.

To add substituents to the main chain, in addition to the list of main chain atoms and attachment points, there is an option to use the hydrogen atoms of the main chain for substituent substitution.

Filters are provided during the formation of the connection between the substituent and the main chain, as well as during the assembly of the main chain. Filters are used to prevent the formation of bonds between oxygen and nitrogen atoms (O-O, N-N, N-O) as well as between oxygen, nitrogen on the one hand and halogens on the other. This makes it possible to remove unstable peroxides, hydrazides, and oximes from consideration.

The Monte Carlo algorithm provides for setting the weights of groups when a random choice of a substitute is made. The probability of choosing a given substituent is equal to its weight divided by the sum of the weights of all substituents. In this version of the program, the weights are the same for all attachment points, but in the future it is planned to make individual weights for each attachment point.

In addition, the functionality of specifying a list of substituents for each atom in the main chain is provided. To do this, atoms are numbered in all selected groups of the main chain. When choosing substituents, you can specify the numbers of atoms in the main chain (Atom Numbers control) to which this substituent is attached (Fig. 5). If the list is defined for at least one of the atoms, then when using the iteration algorithm, the use of attachment points and the Use backbone hydrogens as connection points option are disabled.

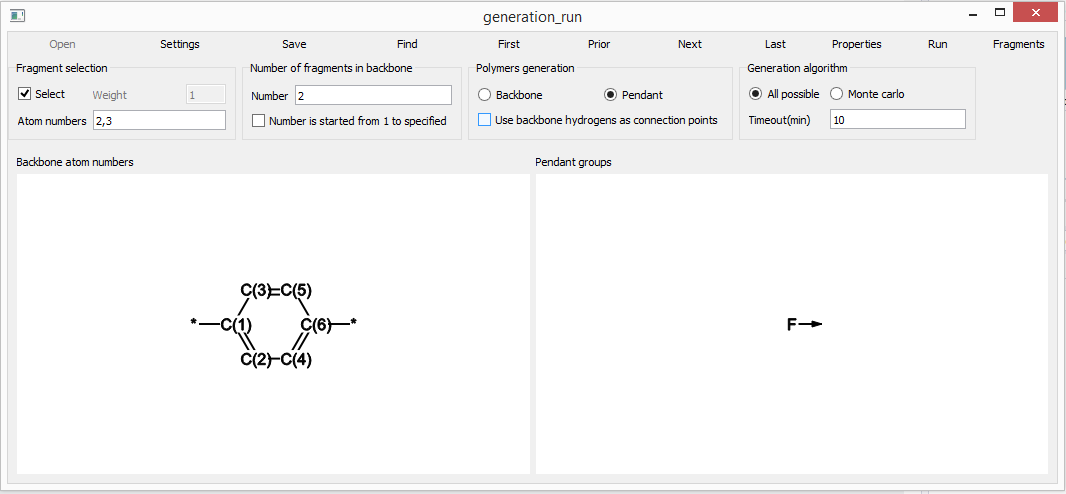


Figure 5. Specifying points of attachment of fluorine atoms (atomic numbers 2,3 in the left part of the figure) to the para-phenylene fragment.

It is difficult to set all these conditions using command line parameters. Therefore, to generate new polymer structures, data is transmitted in an XML file of the format

## File format of initial parameters for generating new chemical structures of polymers.

An example of the contents of the file is shown below:

<?xml version="1.0"?>

<GenerationOptions>

<SDFFragmentFile>C:/kintech/\_polymers/Generate/fragDB/FragNew/fragList.sdf</SDFFragmentFile>

<BackboneFragments>GLUUGHFHXGJENI-RYDMESDDSA-N VNWKTOKETHGBQD-RKEGKUSMSA-N XDTMQSROBMDMFD-KTQOPSMUSA-N</BackboneFragments>

<PendantFragments>UFHFLCQGNIYNRP-OUBTZVSYSA-N VEXZGXHMUGYJMC-OUBTZVSYSA-N</PendantFragments>

<BackboneAtomNumbers>2,3 2,3,4,5</BackboneAtomNumbers>

<NumberBackboneFragments>2</NumberBackboneFragments>

<TimeOutMinutes>10</TimeOutMinutes>

<VariableBackboneLength>true</VariableBackboneLength>

<UseBackboneHydrogensAsConnectionPoints>false</UseBackboneHydrogensAsConnectionPoints>

<UseAlgorithmMonteCarlo>true</UseAlgorithmMonteCarlo>

<PredictedProperties> O2PERMEABILITY H2OPERMEABILITY </PredictedProperties>

<PendantWeights>100 10</PendantWeights>

</GenerationOptions>

Parameter values

<SDFFragmentFile> - path and name of the SD file containing fragments of the main chain and substituents.

<BackboneFragments> and <PendantFragments> are space-separated InChIKeys of the backbone fragments and substituents chosen to generate. These InChIKeys must be in the base SD file.

<BackboneAtomNumbers> - for each selected substituent, contains a list of atom numbers in the main chain fragments to which this substituent is attached.

<NumberBackboneFragments> is the number of backbone fragments used to generate the polymer backbone - two in this case. If the value of <VariableBackboneLength> is false, then a pair of fragments from those listed in <BackboneFragments> are selected in any combination. A value of true means that up to and including two fragments can be used in this case. That is, one or two fragments listed in <BackboneFragments> will be selected to generate the main chain.

<TimeOutMinutes> - task completion time. Must be specified for the Monte Carlo algorithm.

<UseBackboneHydrogensAsConnectionPoints> - if true, then each hydrogen atom attached to the main chain is considered as a substituent attachment point.

<UseAlgorithmMonteCarlo> - if true, then the Monte Carlo algorithm is used, otherwise, all possible variants of polymer structures are enumerated.

<PredictedProperties> - lists the abbreviated names of all properties to be predicted. If you need to predict all available properties, then this section is not in the XML file

<PendantWeights> - integers, proxy weights. They are listed in the same order as they are listed in <PendantFragments>. Works only for the Monte Carlo method.

### Polyprop Program

Console application. Prediction of the properties of a single structure.

Command line options

"structure,s" Chemical structure file MDL \*.mol file format. Atoms with a star symbol (\*) are used as indicators of a repeating polymer fragment.

"result,r" Results file name Machine-readable format of prediction results for further processing by other programs

"ports,p" XML file with results of calculations - XML file with results for Multicomp

"report,e" User-friendly html file with results of calculations - properties prediction results for user display.

Below three parameters are used to calculate the temperature dependence - start temperature, end temperature and number of temperature intervals. If there are no parameters, then the prediction is made for a single temperature of 298K.

"initialtemperature,t" First temperature point, used to calculate temperature dependence

"finaltemperature,f" Last temperature point, used to calculate temperature dependence

"nintervals,n" Number of intervals between temperature points

"molweight,m" Average molweight of polymer. Molecular weight. If not defined then it is considered equal to 100000 Daltons.

"input,i" Input file with initial data Input data from Multicomp in XML format - temperatures for property prediction, molecular weights for property prediction, molecular structure in Multicomp Geometry format or MOL MDL format, PRU length calculated by various methods.

"properties,w" List of properties to be calculated as space-delimited string. See predict\_bicerano.h. Empty string means all properties. Example: -w'O2PERMEABILITY H2OPERMEABILITY' List of properties to predict, separated by spaces. Their description is given above. The absence of this parameter means that all properties will be predicted. Must be enclosed in quotes as spaces on the command line are treated as a new option.

### Genstruc program.

Console application. Used for two purposes

- Creation of a database of chain fragments and substituents

-Generation of new polymer structures and prediction of their properties.

Command line options

"job,j" Job type: create fragment database (=d) generate and predict (=g)

"settings,s" Generation settings The option is active for the job=g parameter. Name and path of the XML file with initial values for generating structures. Its format is described above. Mandatory parameter for polymer generation.

"result,r" Result SD file name Contains the path and name of the SD file. If the job=d parameter, then the database of fragments and substitutes is placed in this file. If the job=g parameter, then the generated polymers and their properties are placed in this file. If this parameter is omitted for the job=d parameter, then the database is written to the fragList.sdf file in the same folder as the molfiles.

"molfolder,m" "Folder with polymer molfiles" Only works for job=d. Path to folder with molfiles. Works with job=d, required parameter

## GenerationRun program

Graphical interface for creating a file with initial data for the Genstruct program (Fig. 6)

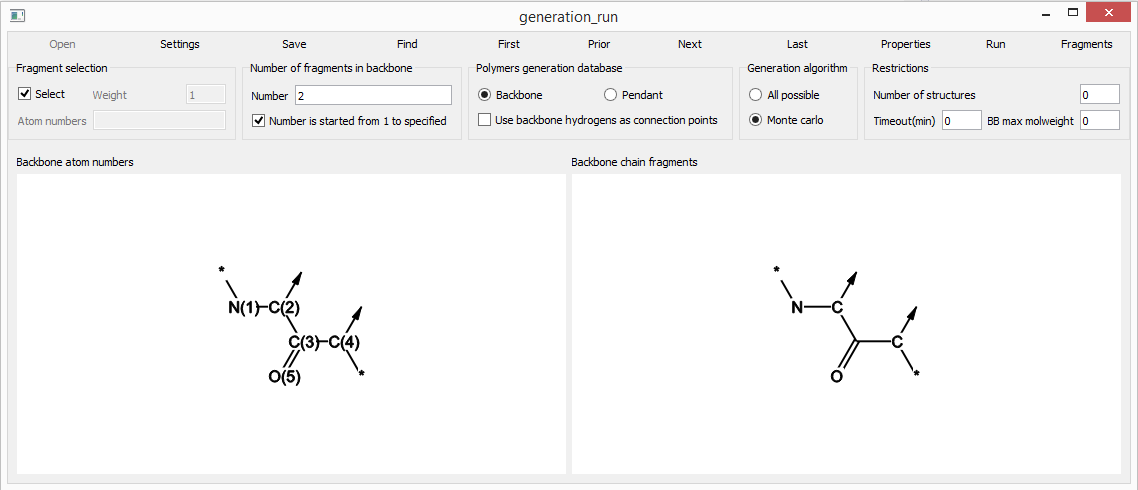


Figure 6. Appearance of the graphical interface for setting the initial conditions for generating structures. The isotopic features of the fragment in the database are interpreted as the sites of attachment of substituents (arrows in the figure) and as designations of the main chain (atoms to which star atoms are attached).

Teams:

Open - Opens an SD file with a database of fragments and substitutes. This file is pre-created by the Genstruc program.

Settings - opens a previously saved settings XML file to generate structures. This file stores the path to the SD file with the base for generating structures, so this command can be used instead of the previously described Open command

Save - saves the current generation settings to an XML file, which is used as an input parameter for generating structures by the Genstruc program. This file can be opened by the previous Settings command

First, Prior, Next, Last are used for navigation in fragment and placeholder databases. Navigation is carried out separately for fragments of the main chain and substituents - depending on the choice in the Polymer Generation control (see below).

### Properties-dialog, allows you to select properties for prediction (Fig. 7)

The meaning of the Select All and Deselect All commands is obvious. OK confirms the selection, Cancel cancels.

It should be noted that the minimum set of properties is selected. If, for example, the prediction of a property requires the value of another property that is not selected in this dialog, Genstruc will automatically add it to the list of predictable properties.

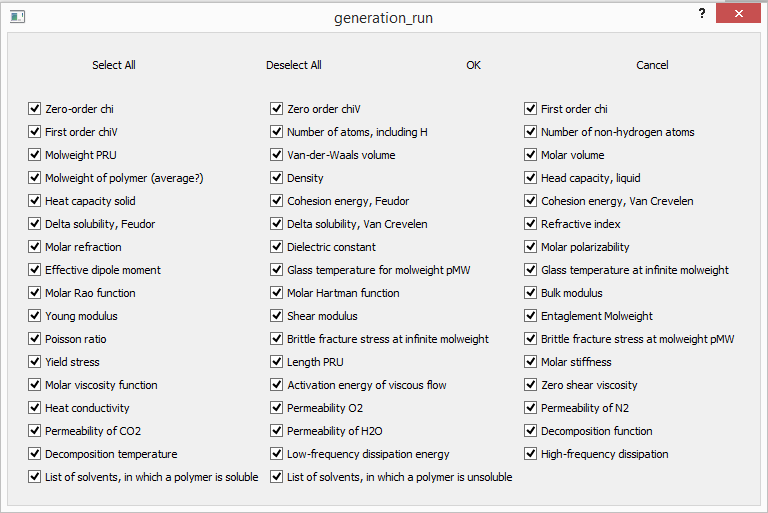


Figure 7. Dialog for selecting properties for prediction.

The two Run and Fragments commands are active only if the same folder as the GenerationRun.exe file also contains the genstruc-d.exe file (the Gensctruc program).

Run - Runs Genstruc to generate new polymer structures using the current parameters. In the dialog that appears, you must specify the name of the SD file to save the results of the prediction-structure of new polymers and their properties

Fragments - Runs Genstruc to create a file to store the fragments and replacements database. In the dialog that appears, you must select the folder with mol files of polymer structures. The SD file containing the base of fragments and substitutes will be saved in the same folder with the name fragList.sdf

### Parameter setting controls

Fragment selection - selects fragments and substitutes for generation (Select checkbox) and also sets the weights of substitutes (Weight, active only for Monte Carlo algorithm).

Atom Numbers - Only works for placeholders. The numbers of atoms in the main chain to which this substituent can be attached are indicated. These numbers are visible in Figure 19, bottom left. If the numbers of atoms are specified for at least one of the substituents, then the points of attachment are ignored, as well as the option of replacing hydrogen atoms in the main chain of the polymer. If lists of atoms are not specified for some substituents, then they are ignored despite the enabled Fragment Selection option

Number of fragments in backbone - sets the number of fragments from the backbone database that will be used to form a new polymer backbone. If the Number is started from 1 to specified control is checked, then no more than the specified number of fragments will be included in the main chain.

Polymer generation - determines for which base fragments are viewed and selected - backbone-base of fragments of the main chain, pendant - base of substituents. In addition, here you can also allow the use of hydrogen atoms that are attached to the main chain as points of attachment. If the Use backbone hydrogens as connection points check box is not checked, then substituents will be added only at connection points (atoms are marked with arrows in chemical structures).

Generation algorithm - choice of algorithm: enumeration (All possible) or random selection of chain fragments and substituents (Monte-Carlo). In addition, the time in minutes before the task is completed is set. 0-time is not limited. It is not recommended to set 0 for the Monte Carlo algorithm - the task will never complete.