Available online at www.sciencedirect.com
ScienceDirect
Mendeleev Commun., 2024, **34**, 792–794
Mendeleev Communications

On evaluating the possibility of synthesizing virtually designed polymers

Sergey V. Trepalin,[a,b] **Pavel V. Komarov,***[c,d] **Andrey A. Knizhnik,**[a,e] **Denis B. Shirabaykin,**[a]
Alexander S. Sinitsa[a,e] **and Boris V. Potapkin**[a,e]

[a] *KintechLab Ltd., 123298 Moscow, Russian Federation*
[b] *Russian Institute for Scientific and Technical Information, Russian Academy of Sciences, 125215 Moscow, Russian Federation*
[c] *A. N. Nesmeyanov Institute of Organoelement Compounds, Russian Academy of Sciences, 119334 Moscow, Russian Federation. E-mail: pv_komarov@mail.ru*
[d] *General Physics Department, Tver State University, 170002 Tver, Russian Federation*
[e] *National Research Center 'Kurchatov Institute', 123182 Moscow, Russian Federation*

DOI: 10.1016/j.mencom.2024.10.008

A method for assessing the synthesis complexity of *in silico* designed polymers using the modified Synthetic Accessibility Score Index is presented. This index is calculated by decomposing a chemical structure into smaller fragments and evaluating their contribution to the overall score. Also discussed are the data sources for these estimates, ways to overcome ambiguity in the representation of the smallest repeating unit of polymers, and the problem of evaluating the stability of chemical structures stored in machine-generated databases.

Keywords: polymers, QSPR, synthesis, chemical databases, structural fragments, polymer stability.

The current state of computer modeling technologies such as big data and machine learning makes it possible to design new polymers virtually (*in silico*). Direct laboratory synthesis can only be used once a list of predicted polymers with the desired properties has been compiled and analyzed. In addition, it is also useful to check the synthesis accessibility and the cost of mass production of selected structures.

The problem of selecting a promising option from a large number of possible polymer structures can be solved using the following computational sequence: 'generation' → 'prediction of properties' → 'selection'. The first and second stages use both traditional methods (combinatorial libraries,[1] virtual chemical reactions[2] and regression models for prediction[3]) and neural networks,[4,5] which is an integral part of the methodology based on AI (Artificial Intelligence). This approach typically requires the creation of databases containing millions of records of new chemical structures, some of which are unstable under normal conditions or difficult to synthesize. Therefore, an important task is to filter the contents of databases in terms of complexity of laboratory synthesis as well as stability. Topological analysis is a useful tool to examine and compare the structures of compounds.[6,7]

When creating the PI1M database (Benchmark Database for Polymer Informatics),[8] the Ertl and Schuffenhauer approach was used to assess the complexity of the synthesis of generated structures.[9,10] It is based on the decomposition of the 2D structural formula of a compound into atom-centered fragments with a topological radius of 3 or less. Their frequency of occurrence is then compared with the frequency of occurrence of fragments in the PubChem database.[11] It is assumed that structures more frequently presented in PubChem can be easily synthesized in the laboratory. Thus, the more fragments with a high frequency of occurrence there are in a new compound, the easier it will be to synthesize. Ertl and Schuffenhauer[9] proposed making such estimates based on calculating the synthesis complexity index SAscore (synthetic accessibility score), which takes into account frequency of occurrence and a number of other parameters. However, this approach is not applicable to polymers, since the polymer backbone continuation symbol (asterisk atom, –*) is considered a fragment of the structure (Figure 1). This conclusion follows from the analysis of SAscore values for polyethylene in the PI1M database[8] (Table 1).

Therefore, the main goals of this work were to develop a new algorithm for assessing the complexity of polymer synthesis and select a method for predicting their stability.

The Ertl and Schuffenhauer algorithm[9] was implemented in the commercial system Pipeline Pilot.[10] However, information about the contributions of different chemical fragments of molecular structures to SAscore is not publicly available. Therefore, in the first stage, a database of fragments of molecular structures was created and their contributions to the SAscore were assessed. For this purpose, as before,[9] an analysis of the PubChem database[11] was carried out. Chemical structures of 91 071 443 of 115 668 082 molecules available in PubChem (as of August 2023) were fragmented using the described method.[9]

Table 1 SAscore values for different polyethylene encodings (file PI1M_v2.cvs[6]).

SMILES	*CC*	*CCC*	*CCCC*
Reference 6	7.58021535416	6.88788163673	6.33350939444
This work	0.41179857709	0.41179857709	0.41179857709

As a result, the generated database contains 7205584284 fragments, of which only 9106672 are unique. The identified 3766913 unique fragments (41%) are singletons, *i.e.*, they occur only once in the dataset under study. It should be noted that Ertl and Schuffenhauer used a representative dataset of 1000000 compounds for fragmentation, in which 51% of singletons were found.[9] This is not surprising since the diversity of chemical structures is not infinite and the percentage of singletons decreases as the size of the dataset increases. The comparison showed that the 23 most frequently occurring fragments in the dataset under study were also included in the published list of the 28 most frequently occurring structures.[9] The remaining fragments are also at the top of the created list, occupying positions 29, 31, 33, 37 and 46 in the overall list of 9106672 fragments.

The occurrence of each chemical structure fragment is expected to be proportional to the size of the dataset. Therefore, to calculate the contribution of the *i*-th fragment to the synthesis complexity value, it was proposed to use the following equation:

$$\mathrm{fragmentScore}(i) = \log[\mathrm{fragOccurrence}(i) \times \mathrm{scale/nFrag80}], \qquad (1)$$

where fragOccurrence(*i*) is the number of occurrences of the *i*-th fragment in the structures, contained in the analyzed dataset, nFrag80 is the number of unique fragments that make up 80% of the dataset.[9] To remove the dependence on the size of the dataset, a scale factor was introduced. It is equal to 1000000/91071443, *i.e.*, the number of structures in the Ertl and Schuffenhauer dataset[9] divided by the number of structures in the training set of this work. Using a large dataset of structures from PubChem, a database of structure fragments with their frequencies of occurrence was created to calculate the SAscore, which is available for public access.[12]

All other factors in the SAscore, such as sizePenalty, stereoComplexityScore and macrocyclePenalty, were determined according to the rules described by Ertl and Schuffenhauer. The calculation results were scaled from a range of −4 to 2.5 to a range of 10 to 1.[9] Note that the authors' approach was not identical to Ertl and Schuffenhauer's due to differences in fragmentScore(*i*), but it was able to produce comparable results. So the final equation for SAscore is as follows:

$$\begin{aligned}\mathrm{SAscore} = 10 - 9[&\Sigma_i \mathrm{fragmentScore}(i) + \mathrm{sizePenalty} + \\ &+ \mathrm{stereoComplexityScore} + \mathrm{macrocyclePenalty} + 4]/6.5.\end{aligned} \qquad (2)$$

It should be noted that in the case of polymers, their chemical structure, usually represented in terms of the smallest repeating units (SRUs), can be encoded in several alternative ways. Several possible ways of representing polyethylene are given in Table 1. It can be seen that they correspond to different SAscore values.[8] To ensure that the analysis results do not depend on the encoding method chosen by the user, the cyclic structure representation can be used as the canonical SRU encoding method, as proposed by Yu.[13] The procedure for generating cyclic SRUs of polymers has been described in detail previously.[1]

A priori, a chemical graph cannot contain links of an atom to itself (loops) and cannot contain multiple links between pairs of atoms. Thus, the SRU of a polymer cannot be represented as a cyclic structure if the topological distance between the backbone continuation points is less than two. In addition, for a cycle size with a topological distance of less than six, the problem of structure fragmentation arises, since atoms in small-sized cycles have to be reincorporated into the description of the fragment. To solve this problem, it is proposed to generate an SAscore value for a cycle consisting of multiple SRUs such that the ring size is at least six (Figure S1, see Online Supplementary Materials). If the topological distance between the repeating backbone atoms is zero (for example, in the case of polymethylene), then a repeating unit consisting of six SRUs is used. For poly(*p*-phenylene) with a topological distance of four between repeating backbone atoms, two SRUs are used, resulting in the formation of an eight-membered ring (see Figure 1). Such transformations make it possible to obtain identical SAscore values for a polymer with any type of encoding.

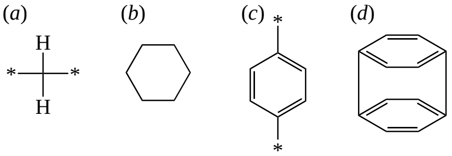

Figure 1 The SRUs of (*a*) polyethylene and (*c*) poly(*p*-phenylene), as well as their associated chemical structures of (*b*) cyclohexane (a cyclic structure derived from polyethylene SRUs) and (*d*) tricyclo[4.2.2.2^{2,5}]dodeca-1(8),2,4,6,9,11-hexaene [derived from poly(*p*-phenylene) SRUs], used to calculate the SAscore of polyethylene and poly(*p*-phenylene). Here the symbol '*' denotes the continuation point in the polymer backbones.

The fragmentScore is normalized to the number of fragments obtained from a chemical structure and, as a result, does not depend on the number of SRUs in the cycle. The sizePenalty and stereoComplexityScore are normalized to the number of SRUs in the cyclic structure and the number of SRUs is calculated by analyzing the topological equivalence of the atoms. These parameters can be considered as the number of atoms in the SRU (sizePenalty) or the number of stereocenters in the SRU (stereoComplexityScore).

The calculated SAscore value for polyethylene [Figure 1(*c*)] is 0.41179 and is identical for all cases included in Table 1 due to the sizePenalty normalization. The cyclohexane SAscore was calculated for *CC* and *CCC*, and the cyclooctane SAscore was calculated for *CCCC* (see Table 1). It is significantly lower than the previously obtained values[8] of 6.5–7.5 (see Table 1). In this regard, the result obtained in this work is considered more realistic since polyethylene is a cheap commercially available polymer. At the same time, polymers with an SAscore higher than seven, according to published results,[9] are classified as difficult-to-synthesize compounds. The reason for this overestimation is the attempt to predict the SAscore index without including the corresponding atom in the neighboring SRU in the fragment instead of the 'asterisk atom'. The simplest way to do this is to transform the SRU encoding into a cyclic representation. The SMILES record[8] contains a continuation of the polymer backbone *C, which, as already written, is not a structural fragment, but, according to the accepted convention, denotes the location of the backbone continuation. If the backbone continuation side is considered a structural fragment, then it is either absent from the original dataset from the PubChem database used in the published work[9] or is rare. In both cases, it is considered a rare fragment with a negative contribution to fragmentScore.

Typically, a new polymer (except those that are easy to recycle) should maintain its performance characteristics over a long period of time. This means that it must be chemically stable. Therefore, it is important to assess the reactivity (stability) of polymers based on analysis of their chemical structure.

The reactivity of polymers was evaluated by the group method, *i.e.*, by searching for predetermined reactive groups (*e.g.*, –OO–) in the polymer structure. For this purpose, a database of reactive groups was created in which they were associated with several possible chemical reactions. Thus, if a reaction involves several reaction groups, a search is made for all corresponding groups in the polymer. The search takes into account the repetition of SRU at the backbone continuation points. For example, a reactive peroxide group will also be found if the polymer structure is specified as in Figure 2.

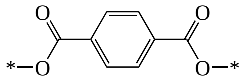

Figure 2 Example of the SRU structure of a hypothetical polymer, poly(terephthaloyl peroxide), containing the peroxide group.

At first glance, it is inappropriate to analyze structures with chemically unstable groups, since they do not occur in practice. However, the development of machine-oriented methods for searching for new polymers based on neural networks often generates compounds with chemically unstable groups. For example, analysis of the PI1M database[8] shows that in chemical structures there, in addition to peroxide groups, there are also other highly reactive groups of atoms with such bonds as halogen–nitrogen and halogen–oxygen.

The database of highly reactive groups currently contains three polymers that are not available for synthesis and 30 highly reactive groups and continues to expand. At the same time, it also includes some reaction conditions, such as temperature, exposure to radiation, *etc.* These additional data are necessary to assess the resistance of polymers to various external factors. For example, the presence of groups that interact with water makes it possible to evaluate the stability of polymers against hydrolysis. In addition to accumulating a database of reactions, further development of the demonstrated algorithm involves abandoning the PubChem database and moving to the generation of structural fragments from databases of commercially available compounds. It should be noted that in addition to stable chemical compounds, PubChem also contains data on compounds that exist at extreme temperatures, such as aluminum monochloride (PubChem CID: 5359282), which is stable in a solid matrix of noble gases at temperatures of 4–20 K or at very high temperatures. There are also compounds that cannot be synthesized in principle, for example nitrogen pentafluoride (PubChem SID: 472285559). Such fragments should be added to the exclusion list that contains information about the impossibility of synthesis. Thus, all detectable chemical structures containing such groups and other highly reactive groups can be easily filtered out at the generation stage.

The algorithm for predicting the complexity of polymer synthesis and the database of reactive fragments have been added to the MULTICOMP software package, developed and supported by the present authors' team.[14] The input parameter is a structure data file (*.sdf) containing the chemical structures of the polymers. The output is the same SD file with an additional field containing the SAscore value and a text field containing the list of reactive groups (if detected). As a check, SAscore values were calculated for the polymers used by Bicerano to parameterize his regression models.[3] The calculation results for the five polymers with the minimum SAscore values and the five polymers with the maximum SAscore values are shown in Table S1 (see Online Supplementary Materials).

The price of a polymer predictably correlates with the complexity of synthesis. Table S1 also shows the announced polymer prices. As can be seen, prices for easily synthesized polymers are generally less than $1 per gram. In the case of polymers with high SAscore values, prices that could be found were over $1 per gram. In most cases, prices were not provided, indicating that such polymers are not commercially available and their laboratory synthesis may not be feasible.

In summary, the known algorithm[9] was extended to assess the complexity of synthesizing periodic polymer structures. Using a large dataset of structures from PubChem, a database of structure fragments with their frequency of occurrence was created to calculate the SAscore, which is publicly available.[12] A relationship was shown between SAscore and market prices for polymers. A mechanism was developed to filter reactive polymers by searching for reactive groups.

P. V. Komarov acknowledges support from the Ministry of Science and Higher Education of the Russian Federation (contract no. 075-00277-24-00).

Online Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi: 10.1016/j.mencom.2024.10.008.

References

1 A. A. Knizhnik, P. V. Komarov, B. V. Potapkin, D. B. Shirabaykin, A. S. Sinitsa and S. V. Trepalin, *Nanomanufacturing*, 2024, **4**, 1; https://doi.org/10.3390/nanomanufacturing4010001.

2 G. Takács, D. Havasi, M. Sándor, Z. Dohánics, G. T. Balogh and R. Kiss, *ACS Med. Chem. Lett.*, 2023, **14**, 1188; https://doi.org/10.1021/acsmedchemlett.3c00146.

3 J. Bicerano, *Prediction of Polymer Properties*, 3rd edn., Marcel Dekker, New York, 2002; https://doi.org/10.1201/9780203910115.

4 C. Yan and G. Li, *Advanced Intelligent Systems*, 2023, **5**, 2200243; https://doi.org/10.1002/aisy.202200243.

5 R. Ma and T. Luo, *J. Chem. Inf. Model.*, 2020, **60**, 4684; https://doi.org/10.1021/acs.jcim.0c00726.

6 I. V. Ananyev and L. L. Fershtat, *Mendeleev Commun.*, 2023, **33**, 806; https://doi.org/10.1016/j.mencom.2023.10.022.

7 E. V. Alexandrov, A. P. Shevchenko, N. A. Nekrasova and V. A. Blatov, *Russ. Chem. Rev.*, 2022, **91**, RCR5032; https://doi.org/10.1070/rcr5032.

8 [dataset] R. Ma and T. Luo, *PI1M: A Benchmark Database for Polymer Informatics*, 2024; https://github.com/ruiminma1996/pi1m.

9 P. Ertl and A. Schuffenhauer, *J. Cheminf.*, 2009, **1**, 8; https://doi.org/10.1186/1758-2946-1-8.

10 [dataset] *BIOVIA Pipeline Pilot*, Dassault Systèmes, 2024; https://www.3ds.com/products/biovia/pipeline-pilot.

11 [dataset] *PubChem*, National Library of Medicine, Bethesda, MD, 2024; https://pubchem.ncbi.nlm.nih.gov/.

12 [dataset] S. V. Trepalin, *SAscore implementation for polymers*, 2024; https://github.com/trepalin/SAscore.

13 M. Yu, Y. Shi, Q. Jia, Q. Wang, Z.-H. Luo, F. Yan and Y.-N. Zhou, *J. Chem. Inf. Model.*, 2023, **63**, 1177; https://doi.org/10.1021/acs.jcim.2c01389.

14 M. A. Akhukov, V. A. Chorkov, A. A. Gavrilov, D. V. Guseva, P. G. Khalatur, A. R. Khokhlov, A. A. Kniznik, P. V. Komarov, M. V. Okun, B. V. Potapkin, V. Yu. Rudyak, D. B. Shirabaykin, A. S. Skomorokhov and S. V. Trepalin, *Comput. Mater. Sci.*, 2023, **216**, 111832; https://doi.org/10.1016/j.commatsci.2022.111832.

Received: 16th April 2024; Com. 24/7470