Chemometrics and Intelligent Laboratory Systems 139 (2014) 121–131

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Prediction of elongation at break for linear polymers

Damián Palomba [a,b], Gustavo E. Vazquez [b,c], Mónica F. Díaz [a,b,*]

[a] Planta Piloto de Ingeniería Química (PLAPIQUI) CONICET-UNS, La Carrindanga km. 7, Bahía Blanca 8000, Argentina
[b] Laboratory for Research and Development in Scientific Computing (LIDeCC), DCIC, UNS, Avenida Alem 1253, Bahía Blanca 8000, Argentina
[c] Universidad Católica del Uruguay, Facultad de Ingeniería y Tecnologías,Av. 8 de Octubre 2801, Montevideo 11600, Uruguay

ARTICLE INFO

Article history:
Received 15 July 2013
Received in revised form 13 August 2014
Accepted 20 September 2014
Available online 26 September 2014

Keywords:
Quantitative structure–property relationships
Molecular modeling
Elongation at break
Mechanical properties
Polymers

ABSTRACT

In this paper we present results on prediction of *elongation at break* (target property) for a group of 77 amorphous polymers of high molecular weight. Novel descriptors are proposed in order to better represent structural features related to the target property. These proposed descriptors along with the classic ones, were calculated for the set of polymers. The final descriptors of the predictive model were obtained by using a combination of variable selection method and domain knowledge. The model consisted of three descriptors: *Cross-head Speed* (*CHS*), *Number Average Molecular Weight/Main Chain Surface Area ratio* (Mn/SA_{MC}), and *Normalized Main Chain Mass* (nM_{MC}). By means of a multi-layer perceptron (MLP) neural network a good prediction model ($R^2 = 0.88$ and $MAE = 1.89$) was achieved, which was internally and externally validated. The model shows the advantages of using well-known parameters in the field of polymers and of capturing the structural characteristics of the main and side chains. Thus, more intelligent tools are developed for the design of new materials with a specific application profile.

1. Introduction

Any engineering activity is dependent on a careful and intelligent selection of materials, including the extraction or preparation of raw products, the design of manufacturing and consumer equipment, the operation and maintenance of a plant, to name but a few. A selection among materials must often be made in order to satisfy requirements of performance and/or cost. In approaching a design problem, the engineer will consider first the desired properties of a specific material. The material performance requirements can be divided into five broad categories, namely functional requirements, processability requirements, cost, reliability, and resistance to service conditions [1]. Although the typical approach in the design of new materials has been empirical (formulation, assembling, synthesis, processing and testing), at present there has been much progress in the knowledge of relationships between the molecular structure of a material and its properties [2]. These advances led to improve the ability to predict the material properties prior to synthesis, which in turn is translated into tremendous savings in time and cost. Nevertheless, it is not easy to achieve these predictions as the variables involved are very complex from a quantitative and quality point of view. Subsequently, the design and synthesis of new materials with specific and novel properties have resulted in one of the most dynamic fields of the modern science [3].

In the materials science, the plastics or polymers are everywhere and their use has been increased almost 20-fold in the last 30 years [3]. Polymers have been modified so as to improve their utility and consequently synthetic polymers were developed. Plastics, fibers, elastomers, adhesives, and coatings have come on the scene as a result of a continual search for man-made substances that can either perform better or be produced at lower cost than natural materials. As a consequence, there was an extraordinary growth in the macromolecule field [4]. In particular, thermoplastics are an interesting set of polymers that become liquid when heated and return to the solid state when cooled. This cycle of melting and freezing can be repeated, so that the plastic can be reshaped by heating it. They are useful for a wide variety of applications, including consumer goods, machine parts, medical equipment, packaging and storage materials. They can be classified as amorphous or semicrystalline plastics, according to their molecular arrangement [5]. Even though amorphous polymers are hard and rigid below the glass transition temperature (T_g), they become soft, flexible and can be shaped above the T_g. Thus, mechanical properties exhibit profound changes in the temperature range where this transition occurs. Semicrystalline polymers have melting points that are above their glass transition temperature. The degree of crystallinity and the morphology of the crystalline phase have an important effect on mechanical properties. Semicrystalline plastics become less rigid above their glass transition temperature yet do not flow until the temperature is above the crystalline melting point. Therefore, when it is attempting to predict the mechanical properties of polymers, it is reasonable to assume that the amorphous polymers behave very differently from semicrystalline ones.

* Corresponding author at: C.C. 717, 8000 Bahía Blanca, Argentina. Tel.: +54 291 4861700x255; fax: +54 291 4861600.
E-mail address: mdiaz@plapiqui.edu.ar (M.F. Díaz).

http://dx.doi.org/10.1016/j.chemolab.2014.09.009
0169-7439/

There are numerous mechanical properties that define the profile of applicability of a polymer and, among them, the ability to resist breaking under tensile stress is one of the most important and widely measured of material properties used in structural applications [6]. For polymers, the tensile test provides vital information related to the ductility and strength through the modulus of elasticity, tensile strength at break and elongation at break, among others. The latest is a measure of material ductility. The *elongation at break* value for brittle materials can be vanishingly small — typically is assumed to be zero. While rigid plastics, especially fiber reinforced ones, often exhibit values fewer than 5%, elastomers and some particularly soft thermoplastics tend to have values above 100%. Materials with higher elongation than 100% have a better capacity to handle an excessive load without failure. Furthermore, not only the rate (cross-head speed), affects the final value of *elongation at break*, but also the ambient temperature does [5]. Values of *elongation at break* reported at specific temperatures and cross-head speeds are typical for these test conditions.

We are interested in studying the possibility of estimating or predicting the mechanical properties of a "virtual polymer" (designed molecule) prior to synthesis. This domain has been little investigated due to its complexity. However, it would be a very useful tool in order to describe the application profile of the new polymer, thereby saving time and cost. Many researchers have adopted computational approaches to predict material behaviors [7]. In particular, the QSPR (quantitative structure–property relationship) technique relates specific parameters of molecule structure (descriptors) to the studied property by using a dataset of molecules and experimental property values. This technique began to be used to predict properties of materials in the late 80s [7]. Since then, the study of materials has been very complex as its properties depend not only on the intrinsic material properties, but also on the history of the material (how it was synthesized, processed, and prepared for testing). Therefore it is important for developing QSPR technique, to generate descriptors that take into account all these aspects; that is, physicochemical property descriptors of materials and descriptors related to the synthesis, processing, or sample preparation to develop the most predictive and useful material property models [7]. Furthermore, by addressing the problem of the synthetic polymers, the difficulty of the molecular design occurs since the depiction of polymeric structures cannot be clearly defined in contrast to the small molecules. Among other factors to consider are: the structural (e.g., chain length, tacticity, and monomer segments) and the composition ones (monomer content, blends, and additives) [8]. For these reasons, it becomes a challenge to generate a reliable associated dataset too. One of the earliest and most widely studied of polymer properties has been the T_g, and good prediction results were obtained from synthetic models [9–24]; in contrast, the mechanical properties of polymers have scarcely been explored. Seitz [25] developed semi-empirical and empirical relationships so as to estimate the mechanical properties of polymeric materials from the molecular weight, van der Waals volume, the length and number of rotational bonds in the repeat unit, besides the T_g of the polymer. Thus, he related the molecular properties of the repeating unit to the properties of the polymer. Ulmer et al. [26] reported the use of a combination of neural networks in the modeling process termed "local property experts" for predicting T_g and other physical and mechanical properties of polymeric materials. The researchers expressed special interest to the design of bisphenol-A polycarbonate (BPAPC) with improved impact resistance. An evaluation of nine BPAPC derivatives by means of the trained neural networks delivered three lead compounds. In a subsequent patent, they claimed that these materials showed improved impact resistance [27]. Eslick and Camarda [28] developed in a preliminary work QSPRs for mechanical properties (tensile strength, elongation at break and 300% modulus) of 35 polyurethane elastomers, using topological descriptors. They utilized a stochastic optimization method to find novel polymers with physical and chemical properties matching a given set of properties for electronic applications. Nevertheless, authors did not present the dataset and detailed explanations of results did not either. A similar work was carried on by Eslick et al. [29], who used computational molecular design (CMD) in cross-linked polymer networks in order to facilitate the development of improved polymethacrylate dental materials. CMD employed QSPRs and optimization techniques to design molecules possessing desired properties, among others tensile strength and modulus of elasticity. The authors used three types of graph to calculate the numerical descriptors (topological) of the polymeric structures: monomer, polymer, and full. Moreover, they computed the degree of conversion and the crosslink density as structural descriptors but any dataset was presented neither for target property nor for descriptors. Holder and Liu [30] developed a quantum mechanically based QSPR model for polymer flexural modulus from structural features of tetrameric oligomers of the polymers. A four-descriptor correlation equation with $R^2 = 0.91$ was achieved using a dataset of 25 polymers. The descriptors in the model showed that rigidity of the monomer, electrostatic interactions and branching were the most important contributors to the flexural modulus value for a particular system. As may be seen, all works had a dataset of very few molecules and most cases did not report the polymer molecular weights and other important structural features for the target property did not either. Nonetheless, these early studies formed the basis to begin the exploration of this research field.

To the best of our knowledge, this is the one of the first attempts to investigate the prediction of tensile properties for polymers by means of the QSPR technique with a reasonable number of consistent and reliable data. In this paper we present results about prediction of *elongation at break* (target property) for a group of polymers. A dataset of 77 molecules was built according to a criterion of common parameters for the tensile test. Simplified molecular models (trimers) were designed so as to depict the polymers and new descriptors were proposed. Then, a combination of variable selection method and domain knowledge was applied to choose the model descriptors. Finally, a QSPR model based on neural networks was developed in order to predict the target property and to provide reliability, interpretability and good performance.

2. Experimental section

In this section experimental aspects are explained in detail according to the usual generation process of a QSPR model: (2.1) Dataset generation, (2.2) Structure entry and optimization, (2.3) Molecular descriptors generation, (2.4) Model development and (2.5) Applicability domain. Below a scheme (Fig. 1) is presented as a guide to the reader with the aim of simplifying and summarizing the entire work.

2.1. Dataset generation

Although our original intention was to work with a dataset from the literature, as usual, we had to deal with the fact that there was none for properties derived from tensile test. Therefore, we began the task of obtaining a reliable and consistent dataset. This was built from information provided by PolyInfo [31]. The dataset polymers and their corresponding observed (experimental) and predicted (calculated) *elongation at break* (%) values are shown in Table 1. Several criteria for selection and creation of dataset were used. Next, a cleaning of dataset was applied. The criteria for selecting, cleaning and a description of dataset are presented below.

2.1.1. Criteria for data selection

The influence of average molecular weights on the behavior of the mechanical properties of polymers is well-known. Furthermore, it is known how important are cross-head speed and temperature in the tensile test on the final value of *elongation at break* of polymers [5]. For these reasons, the following polymer parameters were considered with the aim of building the dataset: number and weight average molecular weight (*Mn* and *Mw*, respectively) and polydispersity index (*PDI*); and as regards tensile test parameters, the following ones

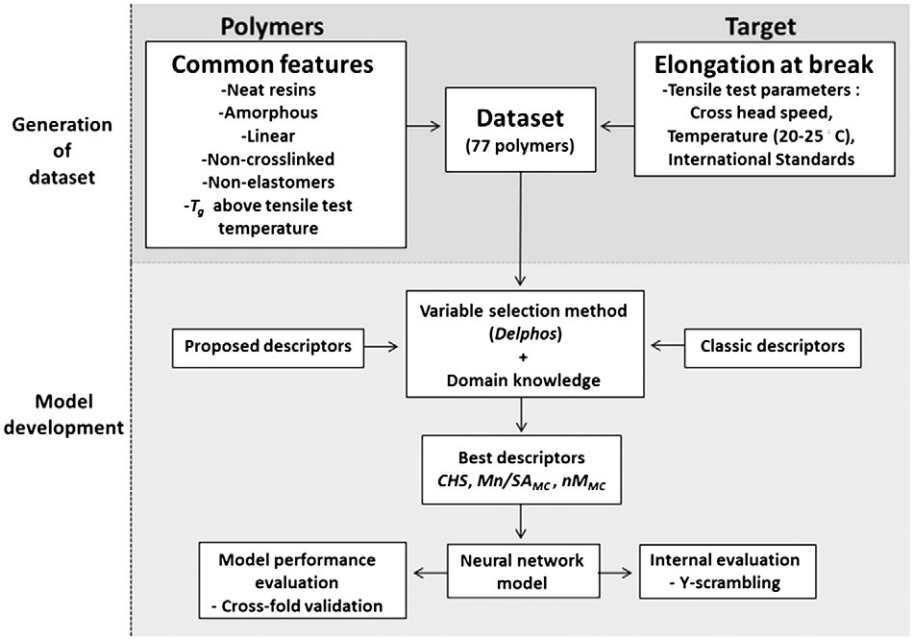

Fig. 1. Scheme of methodology.

were taken into account: cross-head speed and temperature, as well as international standards.

In order to download polymer data from PolyInfo, a script was made. Neat resins (i.e., polymers containing no additives other than an initiation system), average molecular weight and tensile test temperature between 20 and 25 °C were considered. Then, among the polymer data downloaded, only those with *Mn*, *Mw*, and *PDI* data were selected.

2.1.2. Dataset cleaning

All *Mn*, *Mw*, *PDI*, T_g, cross-head speed, and tensile test temperature values for each polymer of the dataset were checked against the original reference. It was confirmed that different values of *elongation at break* from different samples in the same work were due to differences in the average molecular weights of polymers and were not due to another causes, e.g. addition of additives. Likewise, it was corroborated that all dataset polymers have T_g greater than the tensile test temperature (20–25 °C). Furthermore, all semicrystalline, branched, cross-linked, and elastomer polymers were not taken into account, so that the dataset encompassed only amorphous, linear, non-cross-linked and non-elastomer polymers.

2.1.3. Dataset description

The dataset includes: polystyrenes, polyoxides/ethers/acetals, polyesters/thioesters, polyvinyls, polyamides/thioamides, polyimides/thioimides, polyketones/thioketones, polyphenylenes, polysulfides, and polysulfones/sulfoxides/sufonates/sulfonamides. It comprises the following polymer characteristics: linear, thermoplastic, amorphous, flame-retardant, thermally stable, hydrolytically stable, hydrolytically degradable, low toxic, and electroconductive.

The dataset polymer properties have the following ranges of values: *Mn* = 4700–765,000 g/mol, *Mw* = 19,500–2,200,000 g/mol, *Mw*/*Mn* = 1.15–5.6, *Cross-head Speed* = 1–100 mm/min and *elongation at break* = 0.4–39.1%.

Concerning international norms, there are dataset polymers that meet the following ones: ASTM D638, ASTM D882-83, and DIN 53504.53A. With reference to polymerization information, they came from the next types of mechanisms: addition, polycondensation, polyaddition and polymer reaction, polyaddition and polycondensation; and to processing information from: solvent casting, compression, and injection. Finally, regarding shapes of test piece, the following ones were used: film, sheet, dogbone-shaped, and dumb-bell type specimen.

2.2. Structure entry and optimization

Polymers are particularly problematic materials to model since it is clearly not possible to represent (drawing and optimizing) an entire polymer chain in terms of mathematical descriptors [7] because of all polymers' molecular weights are too high. Furthermore, a single molecule is not representative of the whole polymeric material due to the fact that it does not show the weight distribution (typical of these materials). For these reasons, each polymer was modeled using a trimeric structure end-capped by hydrogens where each repeating unit was tail-head bonded. All structures were drawn using HyperChem 8.0.7 [32] and an example is shown in Fig. 2. In turn, two fragments were selected from the middle repeating unit: main chain (MC) and side chain (SC), which were used to generate new descriptors further explained in the next section. The molecules were optimized by using the same software, in order to find energetically stable conformations (those with the lowest energy) that emulate the geometry adopted by a polymer's part owing to intramolecular forces. The structures were optimized with the Force Field Molecular Mechanics (MM+) procedure by using Polak–Ribiere's algorithm and a gradient norm limit of 0.08 kcal/(Å mol).

2.3. Generation of the molecular descriptors

A total of 1041 descriptors were calculated, 51 of them were descriptors proposed by us and the remaining belonged to the classic ones (0D, 1D, and 2D). Then, the whole pool of descriptors was subjected to a combined selection process (domain knowledge and variable selection technique), explained in Section 2.4.1. The most important aspects of the generation of both types of descriptors are described below.

2.3.1. Calculation of proposed descriptors

In a previous paper [33] we introduced descriptors derived from the main and side chains of the middle repeating unit of a trimer (chain descriptors), so as to represent features associated with the structure

Table 1
Dataset polymers including observed (exp.) *elongation at break* (%) and their corresponding predicted (calc.) values.

Sample number	Polymer name	*Elongation at break* (%) (exp.)	*Elongation at break* (%) (calc.)
1	Polystyrene	0.4	0.90
2	Polystyrene	2.04	0.77
3	Polystyrene	1.99	0.90
4	Polystyrene	1.95	0.41
5	Polystyrene	1.4	2.65
6	Polystyrene	1.4	0.51
7	Polystyrene	0.9	1.78
8	Poly((4,4′-oxydianiline)-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})	8	5.68
9	Poly((3,4′-oxydianiline)-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})	5	5.68
10	Poly((4,4′-methylenedianiline)-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})	6	5.81
11	Poly([4,4′-(1,4-phenylenedioxy)dianiline]-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})	7	5.73
12	Poly({4,4′-[1-methylethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})	7	6.48
13	Poly([4,4′-(biphenyl-4,4′-diyldioxy)dianiline]-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})	7	5.52
14	Poly({4,4′-[sulfonylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})	6	6.11
15	Poly({4,4′-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,4-diisopropylbenzene})	7	5.81
16	Poly((m-phenylenediamine)-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})	4	6.47
17	Poly((4,4′-oxydianiline)-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})	6	6.47
18	Poly((3,4′-oxydianiline)-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})	5	6.47
19	Poly((4,4′-methylenedianiline)-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})	6	5.52
20	Poly([4,4′-(1,4-phenylenedioxy)dianiline]-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})	5	5.52
21	Poly({4,4′-[1-methylethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})	6	5.81
22	Poly({4,4′-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{alpha,alpha′-bis[4-(4-carboxyphenoxy)phenyl]-1,3-diisopropylbenzene})	6	5.69
23	Poly({4,4′-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-(5-tert-butylisophthalic acid))	8	9.68
24	Poly{[2-(trifluoromethyl)phenyl]acetylene}	1.7	0.93
25	Poly{[o-(trimethylsilyl)phenyl]acetylene}	4.3	5.51
26	Poly[(2-ethynylphenyl)(trimethyl)germane]	1.6	3.87
27	Poly(1-{2-[(trimethylsilyl)methyl]phenyl}ethene-1,2-diyl)	2.4	5.51
28	Poly[(4-butyl-2,3,5,6-tetrafluorophenyl)acetylene]	8.1	3.51
29	Poly[1-(4-butylphenyl)-2-phenylacetylene]	6.6	1.16
30	Poly{1-phenyl-2-[4-(trimethylsilyl)phenyl]acetylene}	1.5	4.89
31	Poly{1-phenyl-2-[3-(trimethylsilyl)phenyl]acetylene}	2.1	1.29
32	Poly[1-(hexylsulfanyl)prop-1-yne]	20	17.93
33	Poly[1-(decylsulfanyl)prop-1-yne]	21	17.23
34	Poly{(1,1,1,3,3,3-hexafluoro-2,2-diphenyl-propane)-alt-[bis(4-fluorophenyl)methylphosphine oxide]}	26	27.18
35	Poly{hydroquinone-alt-[bis(4-fluorophenyl)methylphosphine oxide]}	38	31.07
36	Poly[(desaminotyrosyl-L-tyrosine hexyl ester)-alt-(succinic acid)]	9	5.52
37	Poly({4,4′-[bis(2-oxodibenzo[c,e][1,2]oxaphoshinin-2-yl)methylene]dianiline}-alt-(terephthalic acid))	3.72	5.81
38	Poly{[4,4′-(9H-fluorene-9,9-diyl)dianiline]-alt-[5,5′-carbonylbis(isobenzofuran-1,3-dione)]}	3.93	1.73
39	Poly{[4,4′-(9H-fluorene-9,9-diyl)dianiline]-alt-[5,5′-carbonylbis(isobenzofuran-1,3-dione)]}	4.91	0.95
40	Poly({O,O′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenylene)]dihydroxylamine}-alt-{5,5′-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	8	11.18
41	Poly({4,4′-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	6	6.43
42	Poly([4,4′-(2-tert-butyl-1,4-phenylenedioxy)dianiline]-alt-{5,5′-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	9	6.48
43	Poly([4,4′-(2,2′-dimethylbiphenyl-4,4′-diyldioxy)dianiline]-alt-{5,5′-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	6	6.77
44	Poly({4,4′-[adamantane-2,2-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	7	6.87
45	Poly({4,4′-[tricyclo[5.2.1.0^2,6^^]decane-8,8-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[4-tert-butylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	7	6.46
46	Poly([4,4′-(1,4-phenylenedioxy)dianiline]-alt-{N,N′-bis[(chloroformyl)methyl]-4,4′-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]diphthalimide})	30.5	28.26
47	Poly({4,4′-[sulfonylbis(4,1-phenyleneoxy)]dianiline}-alt-{N,N′-bis[(chloroformyl)methyl]-4,4′-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]diphthalimide})	39.1	28.99
48	Poly({4,4′-[1-methylethane-1,1-diylbis(4,1-phenyleneoxy)]diphenol}-alt-{N,N′-bis[(chloroformyl)methyl]-4,4′-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]diphthalimide})	29.2	31.00
49	Poly({4,4′-[4-(tert-butyl)cyclohexane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	6.9	6.13
50	Poly({4,4′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	6.5	8.22
51	Poly({4,4′-[(bicyclo[2.2.1]heptane-2,2-diyl)bis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	6.8	6.40
52	Poly([4,4′-(2-tert-butyl-1,4-phenylenedioxy)dianiline]-alt-{5,5′-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	7.6	6.91

Table 1 (*continued*)

Sample number	Polymer name	*Elongation at break* (%) (exp.)	*Elongation at break* (%) (calc.)
53	Poly([4,4′-(2,2′-dimethylbiphenyl-4,4′-diyldioxy)dianiline]-alt-{5,5′-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	7	6.78
54	Poly((4,4′-methylenedianiline)-alt-{5,5′-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	8.5	5.89
55	Poly((4,4′-oxydianiline)-alt-{5,5′-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	7.7	6.35
56	Poly({4,4′-[tricyclo[5.2.1.0^2,6^^]decane-8,8-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[4-phenylcyclohexane-1,1-diylbis(4,1-phenyleneoxy)]diisobenzofuran-1,3-dione})	7.1	6.43
57	Poly({4,4′-[1-(trifluoromethyl)-2,2,2-trifluoroethane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})	7	7.08
58	Poly([4,4′-(2-tert-butyl-1,4-phenylenedioxy)dianiline]-alt-{5,5′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})	6	8.63
59	Poly([2,2′-bis(trifluoromethyl)benzidine]-alt-{5,5′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})	6	10.49
60	Poly([4,4′-(2,2′-dimethylbiphenyl-4,4′-diyldioxy)dianiline]-alt-{5,5′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})	7	8.95
61	Poly({4,4′-[adamantane-2,2-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})	7	7.85
62	Poly({4,4′-[(bicyclo[2.2.1]heptane-2,2-diyl)bis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})	10	7.08
63	Poly({4,4′-[tricyclo[5.2.1.0^2,6^^]decane-8,8-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[1-methylethane-1,1-diylbis(2,6-dimethyl-4,1-phenyleneoxy)]bis(isobenzofuran-1,3-dione)})	10	8.20
64	Poly{(4,4′-oxydianiline)-alt-[6,6′-bis(4-tertbutylphenyl)biphenyl-3,3′,4,4′-tetracarboxylic anhydride]}	9	9.16
65	Poly{(4,4′-methylenedianiline)-alt-[6,6′-bis(4-tertbutylphenyl)biphenyl-3,3′,4,4′-tetracarboxylic anhydride]}	9	10.20
66	Poly({4,4′-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]dianiline}-alt-[6,6′-bis(4-tertbutylphenyl)biphenyl-3,3′,4,4′-tetracarboxylic anhydride])	6	10.11
67	Poly((4,4′-oxydianiline)-alt-{6,6′-bis[4-(trimethylsilyl)phenyl]biphenyl-3,3′,4,4′-tetracarboxylic anhydride})	11	13.47
68	Poly({4,4′-[2,2,2-trifluoro-1-(trifluoromethyl)ethane-1,1-diyl]dianiline}-alt-{6,6′-bis[4-(trimethylsilyl)phenyl]biphenyl-3,3′,4,4′-tetracarboxylic anhydride})	7	13.60
69	Poly([4,4′-(2,2′-dimethylbiphenyl-4,4′-diyldioxy)dianiline]-alt-{5,5′-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})	6	6.78
70	Poly((4,4′-methylenedianiline)-alt-{5,5′-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})	7	5.92
71	Poly([4,4′-(2-tert-butyl-1,4-phenylenedioxy)dianiline]-alt-{5,5′-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})	8	6.48
72	Poly((4,4′-oxydianiline)-alt-{5,5′-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})	10	7.04
73	Poly({4,4′-[(bicyclo[2.2.1]heptane-2,2-diyl)bis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})	9	6.39
74	Poly({4,4′-[4-(tert-butyl)cyclohexane-1,1-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})	7	6.87
75	Poly({4,4′-[adamantane-2,2-diylbis(4,1-phenyleneoxy)]dianiline}-alt-{5,5′-[cyclododecane-1,1-diylbis(4,1-phenylene)]bis(isobenzofuran-1,3-dione)})	6	6.13
76	Poly[sulfonyl(3-sulfo-1,4-phenylene)sulfanediyl-1,4-phenylenesulfanediyl-1,4-phenylenesulfanediyl(2-sulfo-1,4-phenylene)]	16	11.64
77	Poly[sulfonyl(3-sulfo-1,4-phenylene)sulfanediyl-1,4-phenylenesulfanediyl-1,4-phenylenesulfanediyl(2-sulfo-1,4-phenylene)]	12	14.91

of polymers that are known to influence the behavior of the target property. In this regard, this aim is not achieved with classic descriptors.

The advantage of working with a trimer lies in the faster structure optimization and the easier calculation of the descriptors. The trimer segment that best represents the original polymer structure is the middle one (repeating unit), since it is influenced by physicochemical, steric and electronic features of adjacent units, as well as preserves the structural characteristics of the polymer. The main chain was defined as the succession of all atoms (even the hydrogen atoms attached to them) that were in the backbone of the trimer middle repeating unit, and the remaining atoms (in this middle repeating unit) were considered as side chain. Once the molecules were drawn and optimized, the following 7 specific properties were calculated (by using HyperChem) for the main and side chains of the middle repeating unit of the trimer: *Van der Waals surface area*, *Van der Waals volume*, *Log P* (*logarithm octanol–water partition coefficient*), *Refractivity*, *Polarizability*, *Mass*, and *Number of atoms*. To put it another way, these specific properties were estimated for fragments of polymers thus obtaining 14 descriptors, 7 for the main chain and 7 for the side chain. Next, the same specific properties (except for *Number of atoms*) were calculated, but "normalized" via dividing by the atom number of the respective polymer portion considered, bringing another 12 descriptors, 6 for the main chain and 6 for the side chain. Therefore, 26 descriptors were calculated and are detailed in Supplementary file.

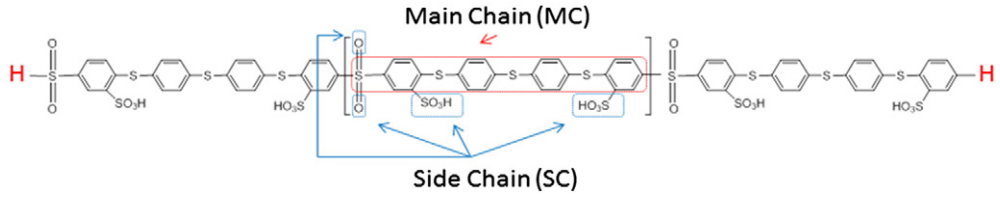

Fig. 2. Trimeric molecular model and identification of its fragments for sample number #76. Example of main chain (MC) and side chain (SC) fragments that belong to the repeating unit of trimeric structure.

Table 2
Nomenclature of proposed descriptors.

New descriptors	Nomenclature	
	Unnormalized	Normalized
Main chain surface area/side chain surface area	SA_{MC}/SA_{SC}	nSA_{MC}/nSA_{SC}
Main chain volume/side chain volume	V_{MC}/V_{SC}	nV_{MC}/nV_{SC}
Main chain *Log P*/side chain *Log P*	$Log\ P_{MC}/Log\ P_{SC}$	$nLog\ P_{MC}/nLog\ P_{SC}$
Main chain refractivity/side chain refractivity	R_{MC}/R_{SC}	nR_{MC}/nR_{SC}
Main chain polarizability/side chain polarizability	P_{MC}/P_{SC}	nP_{MC}/nP_{SC}
Main chain mass/side chain mass	M_{MC}/M_{SC}	nM_{MC}/nM_{SC}
Main chain atom number/side chain atom number	N_{MC}/N_{SC}	–
Number-average molecular weight/monomer molecular weight (number of average repeating units)	Mn/MW	–
Total number of bonds that have bond order greater than one · number of average repeating units	$nBondsM \cdot (Mn/Mw)$	–
Number of rotatable bonds · number of average repeating units	$RBN \cdot (Mn/Mw)$	–
Number of acceptor atoms for H-bonds (N,O,F) · number of average repeating units	$nHAcc \cdot (Mn/Mw)$	–
Number of donor atoms for H-bonds (N and O) · number of average repeating units	$nHDon \cdot (Mn/Mw)$	–
(Main chain surface area/side chain surface area) · number-average molecular weight	$(SA_{MC}/SA_{SC}) \cdot Mn$	–
Number-average molecular weight/main chain surface area	Mn/SA_{MC}	–
Number-average molecular weight/side chain surface area	Mn/SA_{SC}	–

In order to continue developing descriptors for polymers, herein also other chain descriptors have been proposed. These derive from ratio between chain descriptors (calculated for each specific property) and are, in total, 13 new descriptors. Furthermore, other 8 *Mn*-linked descriptors were generated with the aim of combining a priori important information (both macromolecular and repeating unit) for the target property. Thus, it should be noted that many of the proposed descriptors in this paper are related to *Mn* which is an experimental datum. In Table 2, the nomenclature for the 21 new descriptors is shown and their values are available as a Supplementary file.

On the other hand, the experimental parameters *Mn*, *Mw*, *Mw*/*Mn*, and *Cross-head Speed*, which were taken from PolyInfo, were incorporated as descriptors.

2.3.2. Calculation of the classic descriptors

As a supplement of our descriptors, a set of classic variables were calculated by using Dragon 5.5 software [34] and PADEL software [35]. In view of most dataset polymers possess complex molecular structures, these descriptors were calculated on the monomer (not on the trimer) in order to save computational cost. Moreover, these classic descriptors were calculated considering the *whole* monomer (not the user-defined fragments mentioned in Section 2.2), because neither Dragon nor PADEL could select any fragment of the molecule.

Some descriptors were not considered. Besides all binary descriptors, fingerprints (2D binary and 2D frequency fingerprints) [34,36] were not taken into account in order to prevent the introduction of the "missing structure" phenomenon [37]; e.g. the missing structures could be fragments, functional groups, etc. This phenomenon occurs in QSPR models when some fragments do not exist or they have a very low frequency in the training set; hence, the coefficients associated with these sub-structures are not statistically significant [37].

Additionally, some descriptors belonging to the molecular properties category [34,36] were deleted as they are associated to drug features (e.g. drug like index [36]), which are evidently completely different from polymer characteristics. The 3D descriptors [34] were also avoided so as to obtain simpler models. Lastly, constants' descriptors (i.e., variables that take a same value for all samples in the dataset) and near constants (i.e., variables that take a same value, but allowing some predetermined small number of samples to take other values) were excluded.

The final pool of classic descriptors chosen consisted of 990 descriptors and their values are available as a Supplementary file.

2.4. Model development

A variable selection method was included as the first step in the model building process. The purpose of variable selection is to reduce the set of descriptors (independent variables) when predicting a target (*elongation at break*) in order to improve the prediction performance of the descriptors and to provide a better understanding of the underlying process. This technique was applied to the whole set of descriptors (Sections 2.3.1 and 2.3.2). Next, the selected descriptors were used as input for ANNs. A 4-fold cross-validation scheme was employed to evaluate the model performance (see Section 2.4.2 for *k*-fold cross-validation technique explanation). Methodology was summarized in Fig. 1.

2.4.1. Variable selection

When a physical chemistry expert develops a QSPR predictive model, the choice of the most appropriate descriptors for this model constitutes the first complex challenge. Once the molecular descriptors have been computed for a given dataset, different combinations of them should be analyzed in order to obtain a good quality model. In this context, the QSPR model must satisfy two quality standards: high prediction accuracy, statistically evaluated, and good interpretability, evaluated from a physicochemical point of view. Considering the high number of molecular descriptors that are usually calculated for a dataset (sometimes above 1000), a common practice consists of exploring different combinations of descriptors by feature selection methods so as to obtain alternative models. In this work, we adopted a semi-automatic approach with the aim of identifying the most relevant descriptors by a combination of automatic variable selection software and domain experts; a detailed explanation is described below.

In the first stage, we employed the automatic variable selection method; the chosen software for this purpose was Delphos [38]. This tool in particular allows for the identification of linear and nonlinear variables (descriptors), and it was previously applied in the model development for T_g prediction [33]. Delphos brings not only one but also several possible sets of descriptors. In more formal terms, given an arbitrary initial set of descriptors and a target *T*, Delphos produces *p* sets $\{S_1, S_2, \ldots, S_p\}$, so that each S_i $(i = 1, \ldots, p)$ is a "statistically good" set of descriptors in order to establish a QSPR relationship between each S_i and *T*. Moreover, the cardinality of each set of descriptors S_i could be different. Although all S_i subsets are not identical, some descriptors could appear in several subsets; this situation usually occurs with the most relevant descriptors. Delphos is based on a wrapper methodology. It works as follows: In a first phase a multi-objective genetic algorithm is used so as to find the best subsets of descriptors; the fitness function enables different regression techniques to assess these subsets. In a second phase, the best subsets are rigorously evaluated (statistically) by an ensemble of ANN. As a result, Delphos provides as output multiple S_i sets of descriptors best correlated with the target property (*T*), based on the lowest mean absolute error (*MAE*) and mean square error (*MSE*).

Table 3
Numerical values of model descriptors.

Sample number	*CHS* (mm/min)	Mn/SA_{MC} (g mol^{-1} Å^{-2})	nM_{MC} (g mol^{-1})
1	5	4676.45	5.41
2	1.27	5550.91	5.41
3	1.27	29,085.24	5.41
4	1.27	8782.60	5.41
5	5	513.27	5.41
6	5	513.27	5.41
7	5	513.27	5.41
8	50	68.46	8.13
9	50	55.52	8.13
10	50	69.36	7.91
11	50	64.28	8.16
12	50	54.49	8.15
13	50	57.97	8.11
14	50	53.20	8.34
15	50	50.48	8.15
16	50	127.25	8.10
17	50	92.16	8.13
18	50	108.32	8.13
19	50	95.62	7.91
20	50	89.09	8.16
21	50	93.80	8.15
22	50	107.38	8.15
23	5	163.36	8.23
24	30.1	10,744.18	8.34
25	30.1	28,542.07	8.34
26	30.1	10,605.64	8.34
27	30.1	24,449.88	8.34
28	30.1	26,030.13	8.34
29	30.1	57,442.56	12.01
30	30.1	93,691.44	12.01
31	30.1	30,558.61	12.01
32	30.1	4285.89	12.01
33	30.1	4797.64	12.01
34	12.7	64.26	8.62
35	12.7	171.59	8.83
36	100	212.83	7.24
37	50	199.80	7.96
38	1	34.42	8.75
39	1	29.73	8.75
40	20	139.95	9.31
41	20	64.23	8.56
42	20	65.18	8.74
43	20	68.31	8.70
44	20	47.37	8.56
45	20	76.37	8.56
46	9	16.63	8.42
47	9	10.83	8.61
48	9	8.33	8.36
49	20	45.81	8.56
50	20	28.44	8.92
51	20	35.09	8.56
52	20	100.67	8.74
53	20	39.76	8.70
54	20	55.54	8.41
55	20	72.72	8.69
56	20	64.04	8.56
57	20	84.34	8.92
58	20	64.75	9.16
59	20	68.57	9.32
60	20	80.37	9.07
61	20	65.30	8.92
62	20	90.39	8.92
63	20	68.46	8.92
64	5	529.93	9.13
65	5	203.55	8.68
66	5	270.59	9.03
67	5	108.45	9.13
68	5	114.95	9.03
69	20	25.52	8.70
70	20	71.58	8.41
71	20	58.49	8.74
72	20	127.59	8.69
73	20	78.14	8.56
74	20	55.22	8.56
75	20	39.82	8.56
76	5	113.67	10.25
77	5	119.22	10.25

In the second stage of the variable selection of this work, the expert user evaluates the quality of the S_i subsets resulting from Delphos by considering quantitative and qualitative aspects. These S_i should be systematically compared to find the best combination of descriptors. In general, this process encompasses several tasks, such as analyzing descriptor co-occurrence in the different S_i, the relevance of pairwise occurrence of descriptors and descriptor–target relationships. This exploration is carried out by the expert combining the expertise with the design of plots and tables in ad-hoc manner. Thereby, the final selection of the best descriptors, which make up the predictive model, arises from the combination of: the prediction accuracy (minimum prediction error), the physicochemical meaning, the interpretability and the number of selected descriptors.

As a result of application of the methodology explained above, 3 descriptors were chosen by domain experts, who aimed at including into the model orthogonal aspects of the molecules, so that important and interpretable features are considered and redundancy is kept minimal. These selected descriptors are: *Cross-head Speed* (*CHS*), *Number Average Molecular Weight/Main Chain Surface Area ratio* (Mn/SA_{MC}) and *Normalized Main Chain Mass* (nM_{MC}). *CHS* was manually included considering its importance as variable of the tensile test (see details in Section 3.2). Mn/SA_{MC} was chosen due to the fact that it appears in more than one subset and incorporates mass information of polymer molecules. nM_{MC} was the most frequent descriptor among the S_i subsets from Delphos. The descriptors' numerical values are shown in Table 3.

2.4.2. Nonlinear modeling with ANNs

The best set of descriptors obtained as already mentioned in the previous section, was used as input in a multi-layer neural network perceptron (MLP) for the same target (*elongation at break*) by using STATISTICA 8.0 software [39]. The network architecture was defined as MLP 3–3–1 (three input layer neurons, three hidden layer neurons and one output layer neuron); the activation functions were Tanh (Hyperbolic Tangent) for the hidden layer and Logistic (Sigmoid) for the output layer, the error function SOS (sum of squares) and the BFGS (Broyden–Fletcher–Goldfarb–Shanno) quasi-Newton training algorithm.

As for model performance evaluation we employed a *k*-fold cross-validation scheme. It works as follows: the whole dataset is partitioned into *k* equal size subsets; from the *k* subsets, a single subset is retained as the validation data for testing the model, and the remaining $k-1$ subsets are used as training data. In this way, each sample in the dataset is predicted without using it in the training set. In this work *k* is set to 4. The samples selected for each *k* subset were chosen by using a stratified selection. The original list (Table 1) was sorted (ascending) by target values. The compounds designed for validation in each *k*-fold were the following: $\{i^*4+k \mid i=0,1,\ldots; k=1,\ldots,4\}$, and they are detailed in Supplementary material. The reader should note that although we speak in terms of "developing a model using an ANN", the *k*-fold cross-validation scheme requires the training of *k* ANNs.

In addition, Y-randomization technique was applied with the aim of avoiding the possibility of chance correlation of the descriptors. The results are shown in Section 3.1.

2.5. Domain of applicability

As it was mentioned in Section 2.1 (Dataset generation), our dataset was specially assembled in order to have represented different chemical families as well as various molecular weights and testing conditions. Nowadays, the definition of the applicability domain is progressively

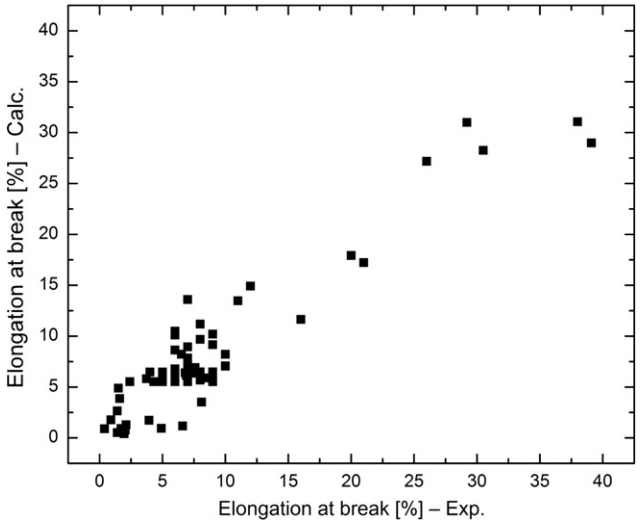

Fig. 3. Calculated vs. experimental values of *elongation at break* (%).

more considered to estimate the reliability of a new prediction (query compound). Following this tendency, in the present work we computed the leverage measurements (extent of similarity). Theoretically, leverage is proportional to Hotelling's T^2 statistic and Mahalanobis distance measure from the centroid of the training set [40]. Usually, a warning threshold is set to three times the average of the leverage p/n, where p is the number of model parameters while n is the number of training compounds. Query compounds with leverage higher than this defined threshold of $3 * p/n$ are considered to be unreliably predicted. For this purpose we applied Leverage node (part of Enalos KNIME Nodes), freely available in the KNIME Community framework [41].

3. Results and discussion

The *elongation at break* is a measure of ductility having a polymeric material. As has already been noted, it is influenced by various factors, including the molecular structure, molecular weight, degree of crystallinity, and testing standards and conditions such as cross-head speed, sample shape, and tensile test temperature. For these reasons, these experimental parameters were considered during the generation of the dataset and we proposed as descriptors some of them with the aim of modeling *elongation at break*. Experimental data collection of many polymers is not a straightforward task due to the fact that most of scientific sources have incomplete and/or inconsistent data. In this fact lies the importance of the dataset presented in this paper, which was defined for amorphous, non-cross-linked, non-elastomer polymers, and whose T_g were above tensile test temperature.

In order to model the target property, the ideal scenario would be to obtain descriptors from a molecular model that represents either the weight distribution of the material or, at least its average molecular weights. Nevertheless, in view of the size and complexity of the entire molecules, it would be impossible to perform this ideal descriptor calculation from a computational perspective [7]. Therefore, a reduced molecular design consisting of a trimer was used to represent each polymer, which proved to be useful in a previous case [33]. The merit of working with a trimer resides in the faster structure optimization and the easier calculation of the descriptors. As mentioned in Section 2.3, the trimer segment that best represents the original polymer structure is the middle one (repeating unit). Thus, it is not surprising that descriptors related to the middle repeating unit (trimer model) were selected in our prediction model and neither of those calculated at the whole monomer.

Even though a trimer is a very simple representation of a polymer, it is valid to optimize its geometry so as to consider its intramolecular interactions. This molecular optimization does not aim to emulate the 3D conformation of polymer molecule, but to consider intramolecular interactions between atoms of neighbor repeating units in minimum scale. Furthermore, the values of the proposed descriptors are affected by these interactions.

The results of this section are presented and discussed as follows: in Section 3.1 the performance of the prediction model, validation and domain of applicability; in Section 3.2 descriptors of the QSPR model and lastly, in Section 3.3 an evaluation of variable relevance.

Table 4
Quality indices for the final model using cross-fold validation.

R^2	*MAE*	*MSE*	*RMSE*
0.88	1.89	6.71	2.59

Table 5
Model acceptability criteria metrics.

Metrics	Fold 1	Fold 2	Fold 3	Fold 4
$R^2 > 0.6$	0.90	0.88	0.89	0.88
$R^2_{ext} > 0.5$	0.88	0.87	0.89	0.87
$(R^2 - R_0^2)/R^2 < 0.1$	0.01	0.00	0.01	0.03
$(R^2 - R'^2_0)/R^2 < 0.1$	0.00	0.01	0.00	0.01
$\lvert R_0^2 - R'^2_0\rvert < 0.3$	0.01	0.01	0.00	0.03
$0.85 \le k \le 1.15$	1.10	0.98	1.02	1.06
$0.85 \le k' \le 1.15$	0.86	0.97	0.93	0.89

3.1. Model performance, validation and domain of applicability

The model performance was assessed using a 4-fold cross-validation technique; the ANNs were developed using the statistical software STATISTICA. The observed and predicted *elongations at break* values are shown in Table 1 and Fig. 3. As can be seen from Table 4, a very good performance is obtained with regard to R^2 (squared correlation coefficient) and other classical statistical parameters.

In view of the dataset range for the *elongation at break* is two orders of magnitude (0.4 for #1 and 39.1 for #47 (from Table 1)). *MAE*, *MSE* and *RMSE* (root mean square error) metrics are not sufficiently representative, mainly for low values. Even so, when a very small value of the target property as for example 0.4 is predicted around 1, from the experimental perspective it would be predicting the same type of material, i.e. a brittle material.

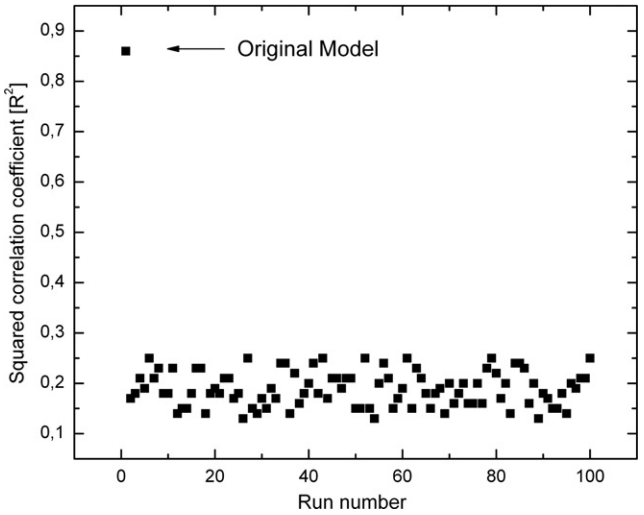

Fig. 4. Y-randomization. R^2 values from 100 models obtained by randomization of the target values (100 runs).

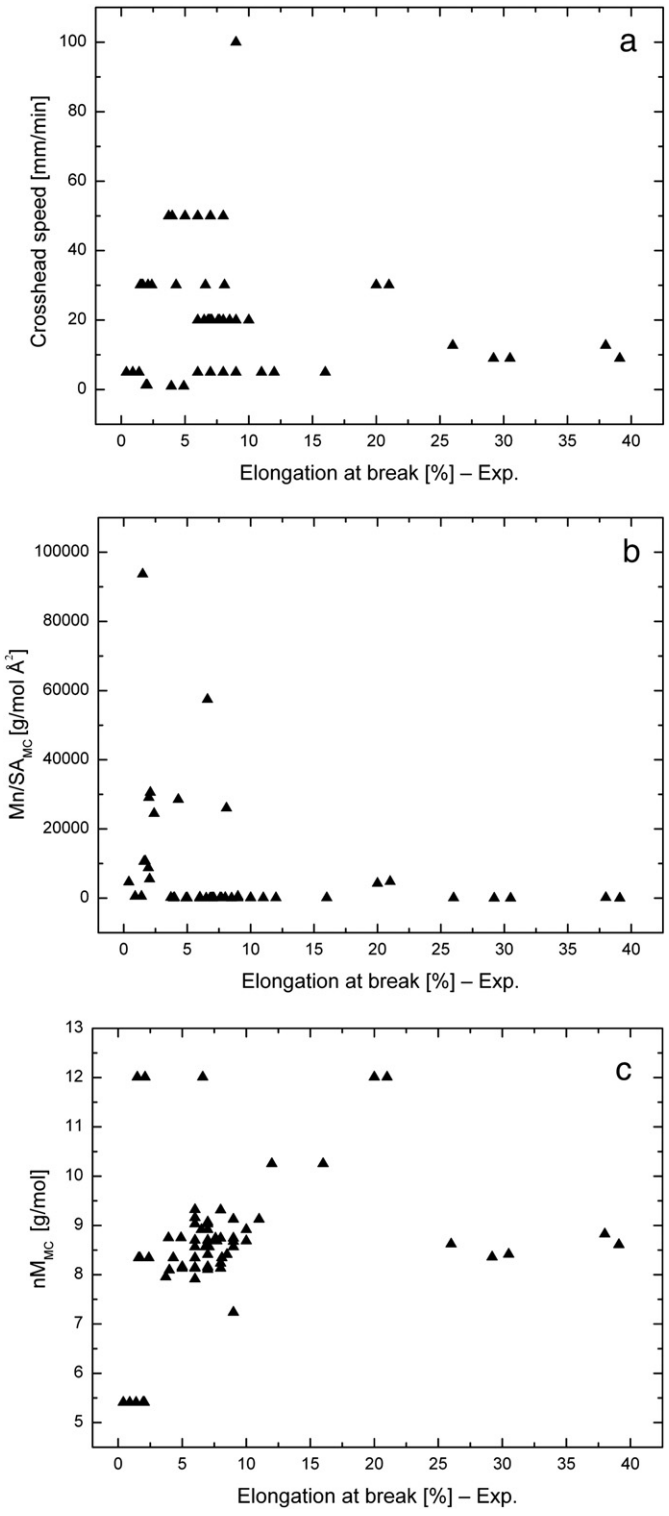

Fig. 5. Plots of model descriptor values versus *elongation at break* values: a) *CHS*, b) Mn/SA_{MC}, and c) nM_{MC}.

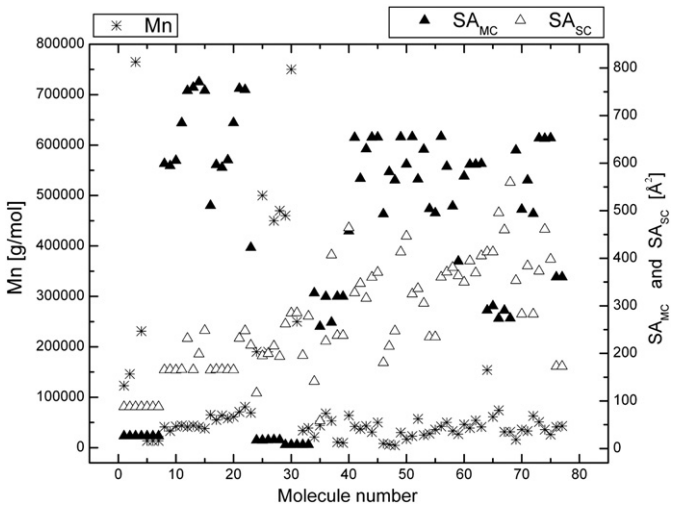

Fig. 6. Plot of *Mn*, SA_{MC}, and SA_{SC} values versus molecule number of our dataset (see Table 1).

Despite the fact that dataset consisted of structurally diverse compounds, only 3 descriptors were used in the model, following the principle of parsimony (Occam) [42]. Hence the generalization ability of model's descriptors is demonstrated.

As already stated above, it is advisable to complete the task with a proper validation. In order to achieve this aim, two different approaches were applied: The model acceptability [40,41] and the internal validation method Y-randomization [43] (also known as Y-random permutation or Y-scrambling).

3.1.1. Model acceptability

According to Tropsha et al. [40], the following statistical criteria must be satisfied by a predictive model: $R^2 > 0.6$, $R^2_{ext} > 0.5$, $(R^2 - R^2_0)/R^2 < 0.1$, $(R^2 - R'^2_0)/R^2 < 0.1$, $|R^2_0 - R'^2_0| < 0.3$, $0.85 \le k \le 1.15$, $0.85 \le k' \le 1.15$; where: R^2 is the correlation coefficient between the predicted and observed activities, R^2_{ext} is the external cross validation, R^2_0 is the coefficient of determination for the predicted versus observed activities, R'^2_0 is the coefficient of determination for the observed versus predicted activities, k is the slope for the predicted versus observed activity regression lines through the origin and k' is the slope for the observed versus predicted activity regression lines through the origin. We computed the above metrics considering as external validation sets the 4 folds previously defined. As expected, all the metrics matched the requirements; results are summarized in Table 5.

3.1.2. Y-randomization

Randomization of target values was applied 100 times; in order to automate the variable selection in each iteration we applied Delphos on the whole set of descriptors and used the best subset that was used to establish a QSPR relationship by using an ANN. The results can be seen in Fig. 4, where all models generated by randomization of target values gave a very poor performance, thereby confirming that there was no chance correlation between the model descriptors and the *elongation at break* values.

3.1.3. Applicability domain

As it was explained in Section 2.5, the applicability domain using leverage measurements was defined for all compounds following a

Table 6
Statistical metrics for the input variable assessment.

	R^2	*MAE*	*MSE*	*RMSE*
Model	0.86	1.88	8.06	2.84
Model − {nM_{MC}}	0.73	2.41	15.44	3.93
Model − {Mn/SA_{MC}}	0.49	2.93	28.37	5.33
Model − {*CHS*}	0.19	3.67	44.48	6.67

LLO (leave-one-out) scheme. The output of applicability domain of each sample was reported in Supplementary file. The results show that only compounds #29, #30 and #36 were estimated as unreliable because this method considers that the predictive values are obtained from extrapolation. However, the corresponding model predictions (Table 1) are still coherent with a brittle material and these chemical families are well represented in the dataset. Therefore, we considered that these three compounds can be included in the model.

3.2. About model descriptors

In the literature for QSPR technique, there is a trend to avoid models that are impossible or very difficult to interpret, although they have good performance. As indicated in [44], "when the interpretation of a QSPR model is consistent with existing theories and knowledge of mechanisms, the ability to explain how and why an estimated value from the model was produced increases. Adding that transparency to model performance is the goal of including a mechanistic interpretation of the model". Even though it is not always possible to find a global interpretation, it is desirable to make the effort to find an explanation for the model in a "mechanistic" way [45]. By following these suggestions, this section is intended to show the type of information provided by each model descriptor in particular and all together in general.

In Fig. 5 the plots of values of the three model descriptors versus the target property values are shown. It can be seen that every descriptor provides supplementary information to the prediction model. Firstly, the *CHS* (*Cross-head Speed*) descriptor is presented in Fig. 5a. In general, in this graph can be observed that the higher the cross-head speed, the lower the *elongation to break values*, regardless of the type of molecule. If, however, low cross-head speed values are analyzed, it can be seen that the target property has a wide spread of values. This situation can be attributed to the fact that *elongation at break*, at low cross-head speed values, becomes independent of this tensile test variable and thus, it can reveal the structure–property relation of each molecule [5] . It can be noted the importance of this tensile test variable, thereby justifying its manual inclusion in the model.

Fig. 5b shows the Mn/SA_{MC} (*Number Average Molecular Weight/Main Chain Surface Area ratio*) values versus the target property values. This descriptor has the special feature of retaining "macro" information of the average real molecule through the value of *Mn*, although its final value is divided by the surface area of the main chain (of the middle repeating unit of the trimer). In order to better understand this descriptor, the *Mn*, SA_{MC} (*Main Chain Surface Area*) and SA_{SC} (*Side Chain Surface Area*) values of the molecules of our dataset are presented in Fig. 6. Note that in our dataset the families of molecules have higher structural variation in the main chain [7.93–770.63 Å^2] than in the side chain [58.32–559.79 Å^2]; therefore, it is reasonable to expect that the descriptors of the model show the higher value spread of our dataset, which occurs for the main chain. In short, this descriptor combines the "macro" information of the molecule (by means of *Mn*) and a structural property related to "micro" information (SA_{MC}).

Finally, Fig. 5c shows nM_{MC} (*Normalized Main Chain Mass*) values vs. target property values. This descriptor provides information about the synthetic model of the polymer, more specifically about the middle repeating unit of the trimer, normalizing the mass of the main chain with its atom number. Once again, the choice of this descriptor by means of the variable selection algorithm could be due to the fact that the information provided by the main chain was prioritized over the information supplied by the side chain.

To sum up, with Fig. 5 it can be concluded that the descriptors are providing information not only on the tensile test, but also on the average real molecule and on the synthetic model segment that has more variation in our dataset. Furthermore, as demonstrated in the following section (3.3), all variables proved to be significant to the performance of the model confirming that the contribution is global.

3.3. Evaluation of variable relevance

Besides statistical measures (Section 3.1), the significance of the input variables to the model was assessed by removing the *i*-th input variable, training the networks without it and evaluating the resulting model by: R^2, *MAE*, *MSE*, and *RMSE*. The metrics were compared with the reference values obtained globally for the complete model (Table 6). When figures showed an important decline, it might be concluded that the presence of the associated *i*-th input variable was compulsory for the model. When figures enhanced or remained similar to original model, the *i*-th input variable should be removed. If the indexes were better, the variable could be affecting the model negatively. In turn, in case they were similar, the variable would seem redundant.

From the results (Table 6), it is worth noting that all the input variables play a fundamental role in the model since none of them neither reach nor overcome original-model performance.

4. Conclusions

In this article we were able to generate a good model of low cardinality (3 descriptors) for predicting *elongation at break* through QSPR approach, within application domain of amorphous polymers.

An original dataset from PolyInfo was generated and presented in this work for the first time, which encompasses neat resins, amorphous, linear, non-cross-linked, and non-elastomer polymers. The difficulty of this comprehensive task resided in the collection of a large amount of consistent experimental data from reliable sources, which did not exist to date in the literature.

New descriptors were proposed in order to better represent structural features related to the target, including experimental parameters. Although all descriptors (new and classic ones) were considered to develop the prediction model, only the proposed descriptors and neither of classic ones were selected by combining a variable selection technique with domain knowledge. This result demonstrated the usefulness of considering a priori important experimental parameters of polymers as descriptors, as well as new structural approaches (main and side chains). Thus, the prediction is tackled from two complementary perspectives.

The prediction model, which was validated by cross-fold validation and Y-randomization, is statistically very good and useful for predicting mechanical properties of polymers provided that certain testing and structural conditions of polymers are met.

This contribution within the framework of the mechanical properties of the polymers allows this research field – totally experimental and applied – to be explored by means of theoretical tools. In this regard, it should be noted that by combining different disciplines (machine learning techniques plus domain knowledge) the reliability of the method was enhanced. Moreover, by using a simplified molecular model (trimer) good prediction results were obtained. Finally, as already has been mentioned, in current literature there are very few studies on prediction of mechanical properties of polymers, which further highlights our endeavor to provide more intelligent tools for the design of new materials with a specific application profile.

Conflict of interest

We certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.

Acknowledgments

Authors thank the National Council of Scientific and Technological Research (CONICET) for supporting this work (Grant PIP 11420110100362) and the SeCyT (UNS) for Grant PGI 24/ZN16.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2014.09.009.

References

[1] Myer Kutz (Ed.), Handbook of Materials Selection, first ed.John Wiley & Sons, New York, United States, 2002.
[2] D.W. Van Krevelen, Properties of Polymers, fourth ed. Elsevier, Amsterdam, The Netherlands, 2009.
[3] L.A. Utracki (Ed.), Polymer Blends Handbook, first ed.Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002.
[4] J. Brandrup, E.H. Immergut, E.A. Grulke (Eds.), Polymer Handbook, fourth ed.John Wiley & Sons, New York, United States, 1999.
[5] W.D. Callister Jr., Materials Science and Engineering: An Introduction, seventh ed. John Wiley & Sons, New York, United States, 2007.
[6] I.M. Ward, J. Sweeney, Mechanical Properties of Solid Polymers, third edition John Wiley & Sons, Chichester, United Kingdom, 2012.
[7] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Quantitative structure–property relationship modeling of diverse materials properties, Chem. Rev. 112 (2012) 2889–2919.
[8] N. Adams, U.S. Schubert, From data to knowledge: chemical data management, data mining, and modeling in polymer science, J. Comb. Chem. 6 (2004) 12–23.
[9] C.Z. Cao, Y.B. Lin, Correlation between the glass transition temperatures and repeating unit structure for high molecular weight polymers, J. Chem. Inf. Comput. Sci. 43 (2003) 643–650.
[10] A. Afantitis, G. Melagraki, K. Makridima, A. Alexandridis, H. Sarimveis, O. Iglessi Markopoulou, Prediction of high weight polymers glass transition temperature using RBF neural networks, Theochem 716 (2005) 193–198.
[11] C. Duce, A. Micheli, A. Starita, M.R. Tiné, R. Solaro, Prediction of polymer properties from their structure by recursive neural networks, Macromol. Rapid Commun. 27 (2006) 711–715.
[12] X. Yu, X. Wang, X. Li, J. Gao, H. Wang, Prediction of glass transition temperatures for polystyrenes by a four-descriptors QSPR model, Macromol. Theory Simul. 15 (2006) 94–99.
[13] J. Gao, X. Wang, X. Li, X. Yu, H. Wang, Prediction of polyamide properties using quantum-chemical methods and BP artificial neural networks, J. Mol. Model. 12 (2006) 513–520.
[14] W.Q. Liu, P.G. Yi, Z.L. Tang, QSPR models for various properties of polymethacrylates based on quantum chemical descriptors, QSAR Comb. Sci. 25 (2006) 936–943.
[15] A. Liu, X. Wang, L. Wang, H. Wang, H.L. Wang, Prediction of dielectric constants and glass transition temperatures of polymers by quantitative structure property relationship, Eur. Polym. J. 43 (2007) 989–995.
[16] X. Yu, B. Yi, X. Wang, Z. Xie, Correlation between the glass transition temperatures and multipole moments for polymers, Chem. Phys. 332 (2007) 115–118.
[17] C. Bertinetto, C. Duce, A. Micheli, R. Solaro, A. Starita, M.R. Tiné, Prediction of the glass transition temperature of (meth)acrylic polymers containing phenyl groups by recursive neural network, Polymer 48 (2007) 7121–7129.
[18] X.L. Yu, B. Yi, X.Y. Wang, Prediction of the glass transition temperatures for polymers with artificial neural network, J. Theor. Comput. Chem. 7 (2008) 953–963.
[19] L.W. Ning, Artificial neural network prediction of glass transition temperature of fluorine-containing polybenzoxazoles, J. Mater. Sci. 44 (2009) 3156–3164.
[20] W. Liu, C. Cao, Artificial neural network prediction of glass transition temperature of polymers, Colloid Polym. Sci. 287 (2009) 811–818.
[21] W. Liu, Prediction of glass transition temperatures of aromatic heterocyclic polyimides using an ANN model, Polym. Eng. Sci. 50 (2010) 1547–1557.
[22] C. Bertinetto, C. Duce, A. Micheli, R. Solaro, M.R. Tiné, QSPR analysis of copolymers by recursive neural networks: prediction of the glass transition temperature of (meth) acrylic random copolymers, Mol. Inf. 29 (2010) 635–643.
[23] I. Hamerton, B.J. Howlin, G. Kamyszek, Predicting glass transition temperatures of polyarylethersulphones using QSPR methods, PLoS One 7 (6) (2012) e38424.
[24] P. Mhlanga, W.A. Wan Hassan, I. Hamerton, B.J. Howlin, Using combined computational techniques to predict the glass transition temperatures of aromatic polybenzoxazines, PLoS One 8 (1) (2013) e53367.
[25] J.T. Seitz, The estimation of mechanical properties of polymers from molecular structure, J. Appl. Polym. Sci. 49 (1993) 1331–1351.
[26] C.W. Ulmer, D.A. Smith, B.G. Sumpter, D.I. Noid, Computational neural networks and the rational design of polymeric materials: the next generation polycarbonates, Comput. Theor. Polym. Sci. 8 (1998) 311–321.
[27] D.A. Smith, C.W. Ulmer, Impact Resistant Polymers. PCT International Patent Application 98/37118, 1998.
[28] J. Eslick, K. Camarda, Polyurethane design using stochastic optimization, 16th European Symposium on Computer Aided Process Engineering and 9th International Symposium on, Process Systems Engineering, 2006, pp. 769–774.
[29] J.C. Eslick, Q. Ye, J. Park, E.M. Topp, P. Spencer, K.V. Camarda, A computational molecular design framework for crosslinked polymer networks, Comput. Chem. Eng. 33 (2009) 954–963.
[30] A.J. Holder, Y. Liu, A quantum mechanical quantitative structure–activity relationship study of the flexural modulus of C, H, O, N-containing polymers, Dent. Mater. 26 (2010) 840–847.
[31] PolyInfo, http://polymer.nims.go.jp/index_en.html (accessed 24.06.13).
[32] HyperChem™, Molecular Modeling System, Release 8.0.7 for Windows, Hypercube, Inc., Gainesville, USA, 2009. (http://www.hyper.com/ (accessed 24.06.13)).
[33] D. Palomba, G.E. Vazquez, M.F. Díaz, Novel descriptors from main and side chains of high-molecular-weight polymers applied to prediction of glass transition temperatures, J. Mol. Graph. Model. 38 (2012) 137–147.
[34] DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.5, Talete srl, Milan, Italy, 2007. (http://www.talete.mi.it/ (accessed 24.06.13)).
[35] C.W. Yap, PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (2011) 1466–1474.
[36] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, second ed. Wiley-VCH, Weinheim, 2009.
[37] I.V. Tetko, P. Bruneau, H.W. Mewes, D.C. Rohrer, G.I. Poda, Can we estimate the accuracy of ADME–Tox predictions? Drug Discov. Today 11 (2006) 700–707.
[38] A.J. Soto, R.L. Cecchini, G.E. Vazquez, I. Ponzoni, Multi-objective feature selection in QSAR using a machine learning approach, QSAR Comb. Sci. 28 (2009) 1509–1523.
[39] STATISTICA (Data Analysis Software System), Version 8.0, StatSoft, Inc., Tulsa, USA, 2007. (http://www.statsoft.com/ (accessed 24.06.13)).
[40] A. Tropsha, P. Gramatica, V.K. Gombar, The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.
[41] G. Melagraki, A. Afantitis, Enalos KNIME nodes: exploring corrosion inhibition of steel in acidic medium, Chemom. Intell. Lab. Syst. 123 (2013) 9–14.
[42] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction, Chem. Rev. 110 (2010) 5714–5789.
[43] C. Rücker, G. Rücker, M. Meringer, Y-randomization and its variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (2007) 2345–2357.
[44] Chapter 6: Guidance on the Principle of Mechanistic Interpretation, Guidance Document on the Validation of (Quantitative) Structure–Activity Relationships [(QSAR)] Models, Series on Testing and Assessment, No. 69OECD Environment Health and Safety Publications, 2007.
[45] P. Gramatica, Chemometric methods and theoretical molecular descriptors in predictive QSAR modeling of the environmental behavior of organic pollutants, in: T. Puzin, J. Leszczynski, M.T.D. Cronin (Eds.), Recent Advances in QSAR Studies: Methods and Applications, Springer, Dordrecht, 2010, pp. 327–366.