Chemometrics and Intelligent Laboratory Systems 191 (2019) 65–72

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Computer-aided design of polymeric materials: Computational study for characterization of databases for prediction of mechanical properties under polydispersity

Fiorella Cravero [a], Santiago A. Schustik [a,b], María Jimena Martínez [c], Carlos D. Barranco [d], Mónica F. Díaz [a,e], Ignacio Ponzoni [c,*]

[a] Planta Piloto de Ingeniería Química (PLAPIQUI), Universidad Nacional del Sur (UNS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Bahía Blanca, Argentina
[b] Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC), Bahía Blanca, Argentina
[c] Instituto de Ciencias e Ingeniería de la Computación (UNS-CONICET), Departamento de Ciencias e Ingeniería de la Computación, (DCIC-UNS), Bahía Blanca, Argentina
[d] Intelligent Data Analysis (DATAi), Division of Computer Science, Pablo de Olavide University, ES-41013, Sevilla, Spain
[e] Departamento de Ingeniería Química (DIQ-UNS), Bahía Blanca, Argentina

ARTICLE INFO

Keywords:
Computer-aided design
Polymeric informatics
QSPR
Feature selection
Artificial intelligence
Material databases

ABSTRACT

In Polymer Informatics, quantitative structure-property relationship (QSPR) modeling is an emerging approach for predicting relevant properties of polymers in the context of computer-aided design of industrial materials. Nevertheless, most QSPR models available in the literature use simplistic computational representations of polymers based on their structural repetitive unit. The aim of this work is to evaluate the effect of this simplification and to analyze new strategies to achieve alternative characterizations that capture the phenomenon of polydispersity. In particular, the experiments reported in this work are focused on three mechanical properties derived from the tensile test. The reported results revealed the disadvantages of using these simplified representations. Besides, we contributed with alternative representations for the databases of polymer molecular descriptors that achieved more realistic and accurate QSPR models.

1. Introduction

The development of machine learning tools for computer-aided design in chemistry is a dynamic area of research [1–4]. In the last decades, among many other artificial intelligence approaches, the use of quantitative structure-property relationship (QSPR) models has increased noticeably [5]. The design of QSPR models constitutes a particular case of predictive modeling problem in which a domain expert is focused on discovering the relationship between some molecular descriptors and a target variable. To infer a QSPR model, it is necessary to identify which descriptors are more related to the target property [6,7]. Molecular descriptors are variables with a key role in characterizing the structure of chemical compounds [8]. Software tools for molecular descriptor computation can calculate thousands of variables but, in general, a regression QSPR model only requires a short number of descriptors for estimating the property under study. Consequently, within the well-known feature selection (FS) problem studied in Computer Science, the selection of descriptors in QSPR modeling is a particular case of it [9].

The computer-aided design and virtual selection of new materials can take advantage of these computational models [10–13]. However, the QSPR modeling in Polymer Informatics is particularly complex [14–17]. In this field, a careful computational modeling of polymeric materials is required. Polymers are huge molecules that consist of a large number of similar structural repetitive units (SRUs) linked together in chain structures [18]. Moreover, a human-made polymer is formed by several chains with different lengths and molecular weights. Therefore, in contrast with a typical drug molecule, a polymeric material is better characterized by a molecular weight distribution curve instead of a single molecular weight value. This is known as polydispersity, and it is a distinctive characteristic of polymeric materials. Considering this characteristic, each of the molecular descriptors of a polymer should be associated with a discrete distribution of values and not with a single value. This distribution is obtained by calculating the molecular descriptor for all chain polymers of

* Corresponding author.
E-mail address: ip@cs.uns.edu.ar (I. Ponzoni).

https://doi.org/10.1016/j.chemolab.2019.06.006
Received 7 March 2019; Received in revised form 18 May 2019; Accepted 15 June 2019
Available online 19 June 2019

different weights [19].

Nevertheless, in general, traditional QSPR approaches proposed for predicting polymer properties in computer-aided design of new materials do not consider polydispersity, oversimplifying the computational representation of each polymeric material to its SRUs [20,21]. In other words, these QSPR models are inferred from datasets in which the molecular descriptors are only computed for the shortest polymeric chain that characterizes each material, without considering neither the remaining polymeric chain lengths nor its associated frequencies (molecular weight distribution curve).

A first preliminary study to consider polydispersity in the prediction of polymer properties is presented in Cravero et al. [22]. In that work, the explored hypothesis is introduced in a general way, and the reported results correspond to a single mechanical property: elongation at break. In addition, the evaluation of the generalizability of QSAR models was performed by calculating an average of the performance values obtained for each of the material representations.

In this work, our main goal is to analyze the impact of this oversimplification in the computational representation of polymeric materials and to evaluate alternative strategies for addressing the molecular descriptor selection problem in a context of polydispersity, considering the molecular weight distribution curve. In particular, we attempt to answer the following questions:

Q1) Is the structural information given by the molecular descriptors related to the SRU-based representation enough for achieving accurate QSPR models?
Q2) Are there any other structural representations of materials based on some characteristic parameters of the molecular weight distribution curves of the materials that yield to predictive models that improve SRU-based models?
Q3) Is it advisable to integrate in a single database the molecular descriptors corresponding to polymeric chains of different characteristic weights related to the molecular weight distribution curves of the materials?

To address these points, we present several databases and QSPR modeling experiments for predicting different mechanical properties of polymeric materials associated with the tensile test. Nonetheless, note that our goal is not the proposal and discussion of QSPR models for these properties. This research is focused on answering the questions related to the structural computational representation of materials previously enumerated to explore a central hypothesis: *QSPR models inferred by using structural information corresponding to several polymeric chain lengths of different characteristic weights of these materials should yield more accurate estimations than QSPR models generated from SRU-based representations.*

The paper is organized as follows: in the next section, the proposed methodology for the analysis is explained. After that, preliminary results are presented for the prediction of three material properties known as *tensile modulus*, *elongation at break*, and *tensile strength at break*, using in-house polymeric material databases. Finally, conclusions and potential further research are discussed.

2. Feature selection in QSPR for Polymer Informatics

The techniques based on QSPR principles estimate a property from molecular, structural, and nonstructural descriptors that numerically quantify different issues of a molecule [14]. In mathematical terms, a QSPR model is defined as a function $Y = f(X)$, where $X = (x_1, x_2, \ldots, x_n)$ is a chemical compound database represented as a vector of molecular descriptors and Y is an experimental target property. The aim is to infer f from a series of chemical compounds whose molecular descriptors have been calculated using HyperChem [23], Padel [24], or other specific tools. Besides, experimental data are also required for the physicochemical property or biological activity of interest (Y). From this database, the function f can be learned by using a training method. Once f has been inferred, it may be applied to new compounds not covered by the training. Thus, f can predict *in silico* the value of a property based on the analysis of data from other experiments. To assess, it is necessary to identify first which molecular descriptors are related to the property under study.

Polymerization is the process in which polymeric materials are created. It consists of the union of several SRUs, which are the minimum part of a polymer. Polymerization produces chains with different lengths. For this reason, polymers are polydisperse. In other words, they have more than one associated molecular weight. A typical curve of molecular weight distribution is presented in Fig. 1 a). The x-axis represents the molecular weight, and the y-axis represents the frequency of occurrence of the chains with each length. Average weights, mainly the weight-average molar mass (Mw) and the number-average molar mass (Mn), are used to describe a polymer. Macromolecule modeling implies building molecular representations with high molecular weights. A computational procedure for *in silico* polymerization is illustrated in Fig. 1 b). Polymer Maker Smiles-based (PolyMas) is a tool developed by our group. This software uses SMILES notation and represents the molecules as strings of character [25,26]. It executes sequential head-tail concatenations of the SRU as many times as necessary to build polymer chains of different lengths.

3. Methods and experimental results

3.1. Target properties and databases

In Polymer Informatics, obtaining material databases is a complex task; moreover if the database must be integrated by polymers related to the study of mechanical properties associated to a tensile test. In this work, an in-house database was developed and applied by our research group [22]. This new database is based on a previous one, also developed by our group [27]. The polymers in this database are homopolymers, linear and amorphous. The 77 initial polymers were characterized by their SRU in SMILES code. To obtain the databases used in this work, it was necessary to polymerize molecules that reached the average weights of the polymers in the database. The Mn varied within a range from 4700 to 765000 [g/mol] and the Mw varied from 19500 to 2200000 [g/mol]. Note that in the context of this work, the term “database” is used to denote the sets of descriptors calculated for each polymeric material representation (SRU, Mn, and Mw); that is, from a single initial database of 77 polymers, different databases emerged depending on the representation used (SRU, Mn, and Mw) for the calculation of the molecular descriptors.

The target properties modeled in this work come from the tensile test. In this test, a polymeric specimen is subjected to a controlled strain (constant cross head speed) until failure. Many mechanical properties are measured during the execution of this test. A typical stress-strain curve for a ductile polymer to define the properties under study is presented in Fig. 2. This curve shows how a material reacts to the forces being applied. In the initial portion of the test, there is a linear relationship between the applied force and the elongation the specimen exhibits. In this linear region, the line obeys the relationship defined as "Hooke's law", in which the ratio of stress to strain is a constant. In other words, the slope of the line in this region where stress is proportional to strain is called the *tensile modulus* or *modulus of elasticity*. On the other hand, the point of failure is more relevant and is typically called breaking or rupture point. The amount of elongation the specimen undergoes during tensile testing is derived from this point; consequently, the ratio of the change in length to the original length is called *elongation at break*. Finally, *tensile strength at break* indicates the stress in this point. In summary, the information about stiffness, ductility, and resistance of the polymeric materials can be obtained from each of these three properties, respectively [28]. This profile of tensile properties can define the application of a material; consequently, it is a key test, and the study of these properties become relevant for new polymeric materials.

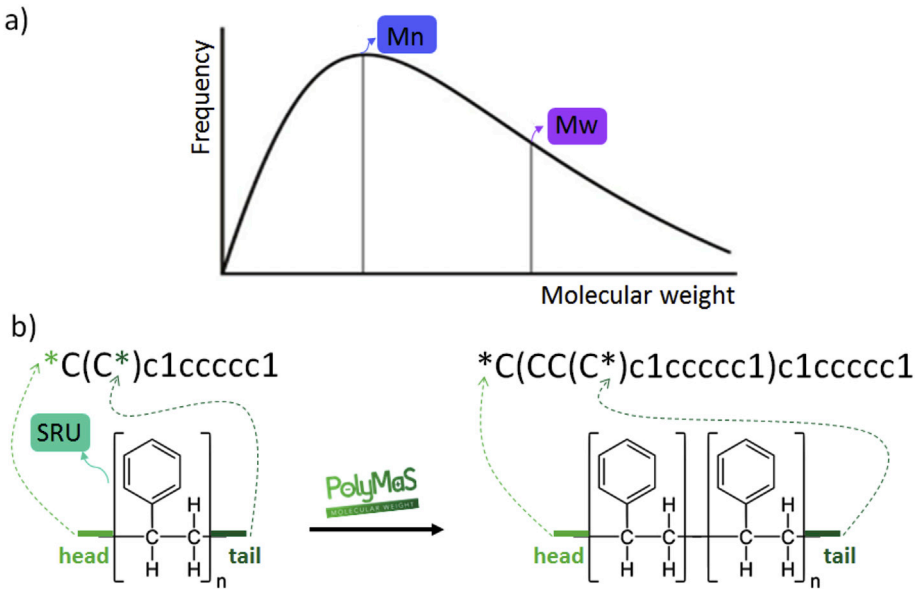

Fig. 1. a) Molecular weight distribution curve. b) Computational procedure for *in silico* polymerization.

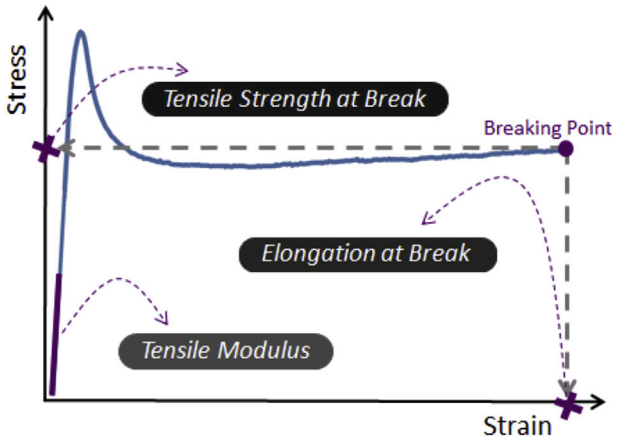

Fig. 2. Stress-strain curve of a ductile polymer.

The proposed experimentation is schematized in Fig. 3. The initial database (DB) consist of 77 polymers represented by their SRU, whose chain length ranges are [5–7205 SRUs] to obtain Mn and [24–10773 SRUs] to obtain Mw. The used representation of SMILES allows us to identify the head and tail of the SRU, that is, the ends, with asterisks (*). We developed PolyMaS with the aim of imitating a polymerization by linking the head of one SRU with the tail of another SRU as many times as the user indicates. PolyMaS joins the head of one SRU with the tail from another SRU, repeating this process until obtaining the desired chain length (see Fig. 1b). Hence, PolyMaS allows obtaining the SMILES codes corresponding to the Mn and Mw weights for each polymer in the DB. Then, each one of these codes together with the SMILES codes from the SRU are the input for the molecular descriptor (MD) calculator algorithm using the RCDK library in R [29]. In this way, molecular descriptors were computed for each material for three polymeric chain lengths corresponding to three characteristic parameters of its molecular weight distribution curve (SRU, Mn, and Mw). Consequently, three different DBs are obtained from the same set of materials.

Although the initial number of polymers was 77 and the maximum number of descriptors calculated was 302, due to limitations of the molecular descriptor calculation algorithms, for some polymers it was not possible to calculate any descriptor, and for other polymers only a reduced set of descriptors was calculated. Therefore, polymers and descriptors were filtered, reducing each DB to 61 polymers and 57 classical molecular descriptors computed by RCDK R library. The names and a brief description of the 57 MD are listed in the Supplementary material. To complete each DB, 51 macro molecular descriptors were added including molecular descriptors with a macro view. These descriptors contain information about testing and real polydisperse material and can be grouped into three classes: 1- the 47 descriptors proposed by Palomba et al. [30] that were calculated on the main chain and the lateral group of the repetitive unit of the polymeric trimer; 2- an important parameter of the tensile test (cross head speed; CHS) [20]; and 3-different average molecular weights (polydispersity index; PDI), average molecular weights in number (Mn), and average molecular weights in weight (Mw) [20,30]. To apply the QSPR model to new molecules, 2- and 3- descriptor values should be added theoretically, according to the desired design for the new material and the tensile test parameters to be performed. This means that PDI, Mn, Mw, and CHS parameters will be estimated values (theoretical) and not experimental ones, because the new material has not yet been synthesized. The list of these 51 macro MD is also included in the Supplementary material.

After analyzing the molecular structure of each polymer in the DB, four polymers (ID15, ID36, ID37, and ID47) were detected as outliers in the external validation set. These polymers were considerably different from the rest of the DB in terms of functional groups and chemical families. Consequently, their representativeness was not guaranteed in the QSPR models trained with other families of polymers. For these reasons, these outliers were eliminated. The correspondence between polymers and chemical families is listed in detail in the Supplementary material.

The three final databases (DB_{SRU}, DB_{Mn} and DB_{Mw}) were integrated by 57 polymers and 108 descriptors. These three databases emerged from the computation of the molecular descriptors for each of the materials representations (SRU, Mn, and Mw). Additionally, once all molecular descriptors for the three databases were computed, joining all the MDs, a global one was defined as DB_{Global}. Therefore, this fourth database contains the information associated to the three different instances of molecular weight of polymeric materials and constitutes a first approach for characterizing polymeric materials by capturing part of their polydispersity.

3.2. Feature selection

The original database was small; however, nine different chemistry

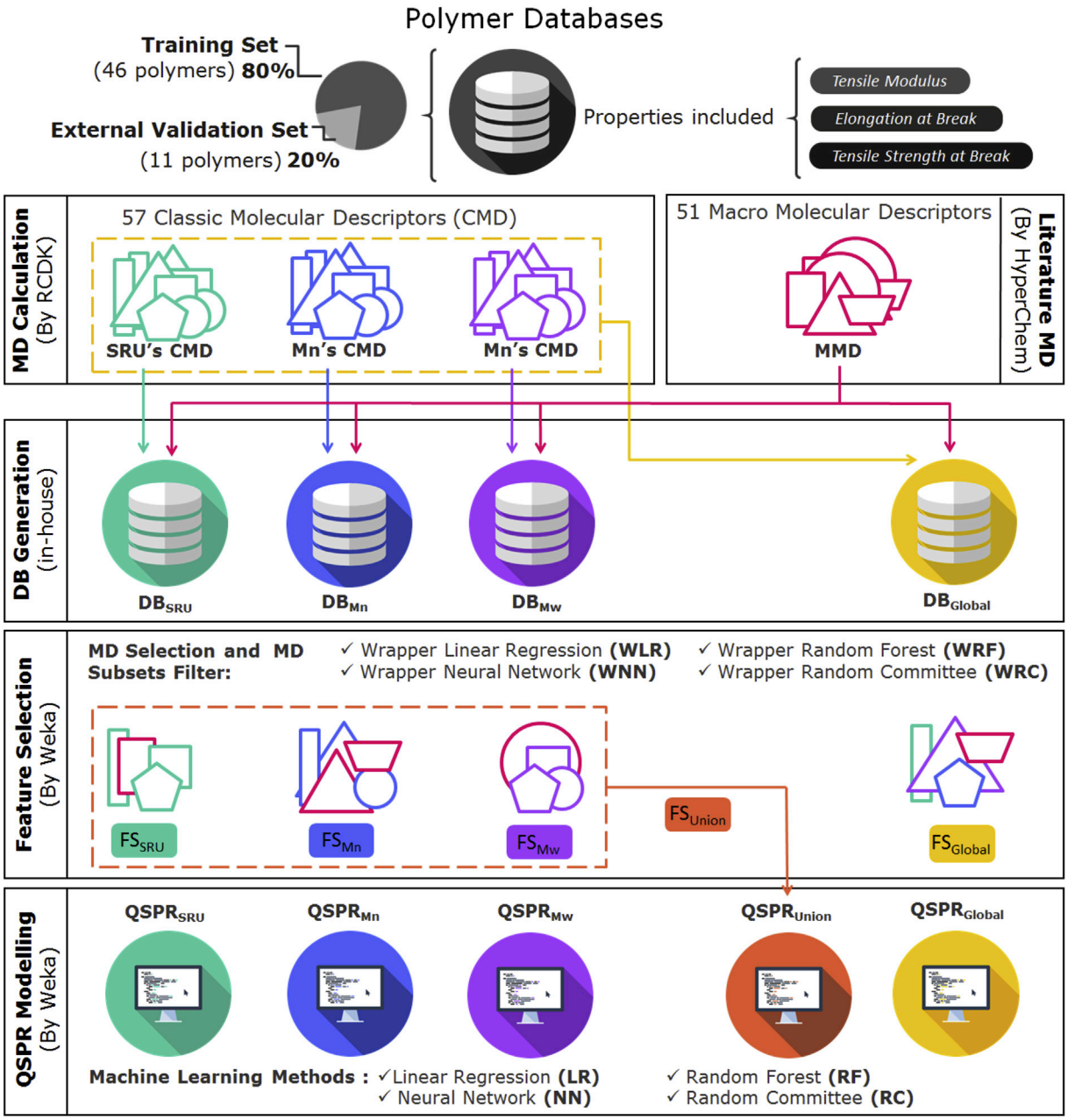

Fig. 3. Scheme of the proposed methodology for the experiments reported in this work.

families were represented there [20]. For this reason, the use of random sampling techniques was not recommended. To ensure the preservation of the global characteristics (complex structural diversity) of the database, it was divided into four folds by a chemistry expert. The complete dataset consisted of molecules that belong to more than one chemical family, because more than one characteristic functional group appeared in their structures. In a first stage, these molecules were grouped by similarity creating a list. Formerly, each member had been chosen in order from this list, from first to last, to fill each of the four folders, balancing the representativeness of the chemical families. For each experiment, databases were divided into two parts. One part consisted of three folds equivalent to ~80% (46 molecules) and was used for the training phase (selection of descriptors and model-making using the leave-one-out cross-validation approach); the other one consisted of one fold equivalent to ~20% (11 molecules) and was used as holdout dataset for external validation during the testing phase (see Fig. 3).

Using the different databases, we designed several experiments to answer the three questions mentioned in the Introduction section of this paper. The WEKA tool [31] was used to select the most relevant molecular descriptors from each database. These feature selection experiments were performed in the training sets using a wrapper method, Best First as search algorithm (W-BF), and four different classification techniques: Linear Regression (LR), Neural Networks (NN), Random Forest (RF), and Random Committee (RC). Once the four subsets of molecular descriptors were selected for each dataset, the domain expert chose the best subsets considering cardinality and balance between the different classes of descriptors. In other words, the QSPR modeler goal was to obtain a subset with few molecular descriptors that belonged to the classical and macro descriptor classes in similar proportions. Following these criteria, in this phase of the analysis we decided to select only two alternative subsets, which will be considered in the final step of the QSPR modeling. Tables 1–3 show the two best subsets of molecular descriptors chosen for each dataset and each property. Note that the MD of all weight instances are present in the DB_{Global}. Therefore, each MD is termed with a suffix

Table 1
Tensile Modulus. Molecular descriptors (MD) for the two best selected subsets for each database (in the training phase). The MDs shared by two or more subsets are highlighted in bold.

DB	FS Method	Classical MDs	Macro MDs	Cardinality
DB_{SRU}	W-BF-RF	**khs.ssNH**, **khs.sssN**, C3SP3, **C4SP3**	nBondsM.(Mn/MW), **nSA_{MC}/nSA_{SC}**, M_{SC}	7
	W-BF-RC	**khs.ssNH**, **khs.sssN**, **khs.sssCH**, nRing4	Mn/MW, nP_{MC}, **nSA_{MC}/nSA_{CS}**	7
DB_{Mn}	W-BF-RF	C2SP3, **C4SP3**, **khs.ssNH**	**nSA_{MC}/nSA_{SC}**	4
	W-BF-LR	**khs.aaaC**, **khs.sssP**	Mn/SA_{MC}, **nP_{MC}/nP_{SC}**, V_{MC}	5
DB_{Mw}	W-BF-NN	C2SP2, khs.aaO, khs.ddssS, **khs.dsssP**, **khs.ssNH**, nAromRings, nRings6	nLogP_{SC}, **nP_{MC}/nP_{SC}**, nP_{SC}, PDI	11
	W-BF-RF	khs.aaCH, khs.aaO, **khs.ssNH**, **khs.sssCH**, **khs.sssN**, nAcid	**M_{SC}**, **nSA_{MC}/nSA_{SC}**	8
DB_{Global}	W-BF-NN	Kier3_SRU, **nAromBond_Mn**, **nAromBond_Mw**, nsmallRings_Mn, tpsaEfficiency_Mw, tpsaEfficiency_SRU	**nSA_{MC}/nSA_{SC}**	7
	W-BF-RF	**C4SP3_SRU**, khs.ssCH2_Mn, khs.ssNH_SRU, **khs.sssN_Mn**	**nSA_{MC}/nSA_{SC}**	5

Table 2
Elongation at Break. Molecular descriptors (MD) for the two best selected subsets for each database (in the training phase). The MDs shared by two or more subsets are highlighted in bold.

DB	FS Method	Classical MDs	Macro MDs	Cardinality
DB_{SRU}	W-BF-NN	**ALogP**, **C2SP2**, nRing4	**nR_{MC}/nR_{SC}**, nSA_{SC}	5
	W-BF-LR	**nRing5**, Zagreb, kier3, khs.dsssP, khs.aasC	**CHS**, LogP_{SC}, nSA_{MC}, **nSA_{SC}**, nR_{SC}, nV_{SC}, PDI	12
DB_{Mn}	W-BF-NN	khs.dssC, **nRing5**	nLogP_{MC}, **nSA_{SC}**, **nR_{MC}/nR_{SC}**	5
	W-BF-RF	**khs.aaO**, **khs.sssCH**, **nRing7**	**CHS**	4
DB_{Mw}	W-BF-LR	nAtomP, **Khs.aaO**, khs.dsCH, VAdjMat	AMR, LogP_{MC}/LogP_{SC}, nP_{MC}/nP_{SC}, **nR_{MC}/nR_{SC}**	8
	W-BF-RF	**khs.aaO**, **khs.sssCH**, **nRing7**	**CHS**	4
DB_{Global}	W-BF-NN	**ALogP_SRU**, **C2SP2_SRU**, **khs.aaO_Mw**	**nSA_{SC}**, **nR_{MC}/nR_{SC}**	5
	W-BF-RF	**khs.aaO_Mw**, khs.ddssS_Mn, **khs.sssCH_SRU**, khs.ssS_SRU, nAcid_Mw, **nRings4_SRU**, **nRings7_Mw**	**CHS**	8

that identifies the weight instance from which it comes from. For example, nAcid_Mn is the MD nAcid (number of acidic groups) calculated for the Mn instance of weight, whereas nAcid_Mw is the same MD but calculated for the Mw instance of weight. Descriptors selected for more than one weight instance are highlighted in bold.

Tables 1–3 show 24 subsets with cardinality ranges 4–11; 4–12; and 5–9 for *tensile modulus*, *elongation at break* and *tensile strength at break,* respectively. Additionally, some MDs seem to be especially chosen in most of the subsets, in particular CHS. This is an important test parameter

Table 3
Tensile Strength at Break. Molecular descriptors (MDs) for the two best selected subsets for each database (in the training phase). The MDs shared by two or more subsets are highlighted in bold.

DB	FS Method	Classical MDs	Macro MDs	Cardinality
DB_{SRU}	W-BF-NN	**khs.dsssP**, **MW**, **nSmallRings**	**nV_{MC}**, **nSA_{MC}/nSA_{SC}**	5
	W-BF-RC	Khs.ddssS, nAromBond, **nAromRings**	LogP_{MC}, **nP_{MC}**, nR_{MC}	6
DB_{Mn}	W-BF-LR	C1SP3, **khs.aaO**, **khs.ssS**	**nM_{MC}**, LogP_{MC}/LogP_{SC}, PDI	6
	W-BF-NN	C1SP2**, khs.dO**, **khs.ssS**, **nAcid**, **nRings5**, tpsaEfficecincy	nSA_{MS}, **nV_{MC}**, PDI	9
DB_{Mw}	W-BF-LR	**khs.aaO**, **khs.ssS**, khs.ssssC, **nAromRings**	**nM_{MC}**, **nSA_{MC}/nSA_{SC}**, P_{MC}	7
	W-BF-NN	**khs.dO**, **khs.sssP**, **khs.ssS**, **nRings5**	**nSA_{MC}/nSA_{SC}**, **nP_{MC}**	6
DB_{Global}	W-BF-NN	**nSmallRings_SRU**, nRings4_SRU, nRings6_SRU, **MW_SRU**, VAdjMat_SRU, khs.aaaC_SRU, **khs.dsssP_SRU**, khs.ssNH_Mn, **khs.aaO_Mw**	-	9
	W-BF-RF	C3SP3_SRU, khs.sCH3_Mn, khs.ssNH_Mw, **khs.ssS_Mn**, **khs.ssS_Mw**, khs.sssCH_SRU, khs.sssN_Mn**, nAcid_Mw**	**nP_{MC}**	9

as it strongly affects the three properties; polymers show very different behavior depending on the testing speed. Another prevalent MD is nSA_{MC}/nSA_{SC}, which represents the molecule size as a ratio of the main and the side chains. Finally, PDI keeps macro information related to the molecular weight distribution. However, many MDs appear for each property specifically. These 24 subsets were used in following training stage.

3.3. QSPR modeling

After the selection of the best molecular descriptor subsets (eight), QSPR models were inferred from each one by using the same four regression methods mentioned before (LR, NN, RF, and RC). Therefore, eight models for each DB: DB_{SRU}, DB_{Mn}, DB_{Mw} and DB_{Global} (i.e., 32 models) were obtained. After assessing these 32 models, only one subset was chosen for each DB, considering the best R^2 and statistical errors, obtaining four final subsets: FS_{SRU}, FS_{Mn}, FS_{Mw} and FS_{Global} (see Supplementary material). Briefly, four final subsets were trained with the LR, NN, RF, and RC learning methods, obtaining 16 models. These models were assessed following the same statistical criteria. Finally, only one model of each final subset was selected: $QSPR_{SRU}$; $QSPR_{Mn}$; $QSPR_{Mw}$; and $QSPR_{Global}$. This modeling process was repeated for each property.

The correlation coefficient (R^2 value), the error metrics of $QSPR_{SRU}$, $QSPR_{Mn}$, $QSPR_{Mw}$, and $QSPR_{Global}$, and the regression method that achieves this accuracy for the external validation set are denoted in Table 4 for all properties. From these results, it is possible to conclude that the SRU-based representation does not guarantee a high performance in all cases. In particular, the $QSPR_{SRU}$ model inferred for predicting *tensile strength at break* is clearly overcome for the QSPR models obtained using higher representations. Therefore, returning to our question Q1 (*Is the structural information given by the molecular descriptors related to the SRU-based representation enough for achieving accurate QSPR models?),* it is possible to answer that, in some cases, the SRU-based representation cannot be enough for inferring accurate QSPR models.

In a similar way, it is possible to conclude that Mn- and Mw-based

Table 4
Cardinality, Machine Learning (ML) method, Feature Selection (FS) method, R^2, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) for QSPR models corresponding to: *Tensile Modulus*, *Elongation at Break* and *Tensile Strength at Break*. For each mechanical property, the two QSPR models that achieved the highest R^2 values are highlighted in bold.

Property	Model	Cardinality	FS Method	Learning Method	External Validation Set		
					R^2	MAE	RMSE
Tensile Modulus	$QSPR_{SRU}$	7	W-BF-RC	RF	0.9679	0.1720	0.2026
	$QSPR_{Mn}$	5	W-BF-LR	NN	0.9629	0.1409	0.1806
	$QSPR_{Mw}$	8	W-BF-RF	RF	**0.9700**	0.1576	0.2102
	$QSPR_{Global}$	7	W-BF-NN	RC	**0.9725**	0.1583	0.1801
Elongation at Break	**$QSPR_{SRU}$**	5	W-BF-NN	NN	**0.8330**	1.0497	1.2064
	$QSPR_{Mn}$	5	W-BF-NN	NN	0.6899	1.3693	1.7031
	$QSPR_{Mw}$	8	W-BF-LR	RF	0.7509	1.0524	1.3414
	$QSPR_{Global}$	5	W-BF-NN	NN	**0.8008**	1.4625	1.6278
Tensile strength at Break	$QSPR_{SRU}$	5	W-BF-NN	RF	0.8377	8.3722	10.6909
	$QSPR_{Mn}$	6	W-BF-LR	RC	0.9267	6.0900	7.2808
	$QSPR_{Mw}$	7	W-BF-LR	RC	**0.9386**	6.0861	6.6046
	$QSPR_{Global}$	9	W-BF-NN	NN	**0.9370**	8.1582	9.4827

representations are clearly overpowered by the $QSPR_{SRU}$ model in the prediction of *elongation at break*. Consequently, regarding our question Q2 (*Are there any other structural representations of materials based on some characteristic parameters of the molecular weight distribution curves of the materials that yield to predictive models that improve SRU-based models*?), it is possible to answer that none of the alternative representations proposed in this paper, based on using Mn and Mw values, reached better performances than the ones achieved by the SRU-based representation for all case studies.

From the results with DB_{Global} (Table 4), we conclude that $QSPR_{Global}$ models achieve a competitive performance in terms of R^2 for the three mechanical properties, supporting the hypothesis that QSPR models inferred using information from several polymeric chains of different characteristic weights of the materials can yield more accurate estimations of the mechanical property values. Note that the $QSPR_{SRU}$ model achieved a better performance than the $QSPR_{Global}$ for *elongation at break*, even when the DB_{Global} includes all MDs of DB_{SRU}. This result can be explained considering that the molecular descriptor selection problem is a particular case of the feature selection problem and is an NP-hard problem in terms of algorithmic complexity. For this reason, any combinatorial optimization procedure for selecting MD subsets can only ensure suboptimal selections. This observation is also valid for the results presented for *tensile strength at break* when the performances of $QSPR_{Global}$ and $QSPR_{Mw}$ are contrasted.

A new subset called FS_{Union} was created by the combination of all the descriptors for the three final subsets belonging to FS_{SRU}, FS_{Mn}, and FS_{Mw} (Fig. 3). Then, this subset was used in the training phase to infer the QSPR models. The objective was to capture features of the three weight instances that compete with the $FS_{Global,}$ which also contained information about all weights. The analysis of the results achieved for the $QSPR_{Global}$ and the $QSPR_{Union}$ can be useful for answering question Q3 (*Is it advisable to integrate, in a single database, molecular descriptors corresponding to polymeric chains of different characteristic weights related to the molecular weight distribution curves of the materials*?). As it was previously explained, in the global database the whole structural information related to the three instances of representation of a polymer material (SRU, Mn, and Mw) were integrated and two different feature selection strategies to choose a molecular descriptor were performed. The first strategy, called FS_{Global}, executes a new feature selection procedure to select the subset of molecular descriptors from the complete set of molecular descriptors included in DB_{Global}. In contrast, in the second strategy, called FS_{Union}, there is not a feature selection procedure, but a union of descriptors from FS_{SRU}, FS_{Mn}, and FS_{Mw} avoiding repetitions. These methodological steps are included in Fig. 3.

The generalizability of the $QSPR_{Global}$ and $QSPR_{Union}$ models can be compared with the unified generalizability quantification of QSPR models generated from unique instances of representation (DB_{SRU}, DB_{Mn}, and DB_{Mw}). This methodology consists of metrics calculation (R^2 and errors) of the alternatives of unique weight instance representations (SRU, Mn, and Mw) in an aggregated manner called $QSPR_{AWI}$ (all weight instances; AWI). For example, the R^2 value of the $QSPR_{AWI}$ model for the *tensile modulus* property is 0.9609. This value was obtained by computing the R^2 value that corresponds to the complete set of prediction results obtained by $QSPR_{SRU}$, $QSPR_{Mn}$, and $QSPR_{Mw}$ for the same property when these models are applied in their external validation datasets respectively (adding the three testing outputs in a unique set of results). Therefore, the validation outputs obtained by these three QSPR models are managed and interpreted as the R^2 of a unified QSPR model. The same procedure is applied for computing the remaining metrics of the $QSPR_{AWI}$ models. Contrasting the accuracies (R^2 values) of $QSPR_{Union}$ and $QSPR_{Global}$, with the accuracies of the $QSPR_{AWI}$ reported in Fig. 4 for the three target properties, it is clear that $QSPR_{Global}$ and $QSPR_{Union}$ outperform the performance of the QSPR models learned from polymer databases corresponding to only one weight instance. Therefore, it can be concluded that models inferred from several weight instances have better generalizability properties.

Regarding $QSPR_{Union}$, its performance is high for *tensile modulus*, but it is slightly good for *tensile strength at break* and poor for *elongation at break*. At first sight, this last result can be unexpected considering that $QSPR_{Union}$ models use more molecular descriptors (i.e., more information) than $QSPR_{Global}$ models (see cardinalities of the models in Table 4). Nevertheless, the union of the molecular descriptor subsets selected from DB_{SRU}, DB_{Mn}, and DB_{Mw} could be combining redundant information, deriving in an overfitting and lacking of generalizability of the $QSPR_{Union}$ models. The generalizability of a QSPR model is another key issue in the performance analysis, which studies the capability to make accurate predictions from unknown data. In other words, this issue is related to the size of the *chemical subspaces* (*subsets of structurally similar molecules*) where a QSPR model can make accurate predictions.

Finally, it is possible to answer question Q3 by concluding that the integration of structural information that corresponds to polymeric chains of different characteristic weights related to the molecular weight distribution curves of the materials is an advisable practice for QSPR modeling in Polymer Informatics. This representation strategy may partially capture the polydispersity inherent to these synthetic materials, benefiting the accuracy and generalizability of the QSPR models without demanding a significant higher number of selected molecular descriptors (*model cardinality*) than the one required for SRU-based models.

4. Conclusions

In Polymer Informatics, an emerging subfield of Cheminformatics, the inference of QSPR models constitute a relevant topic associated with the computer-aided design of new industrial materials. A complex issue for

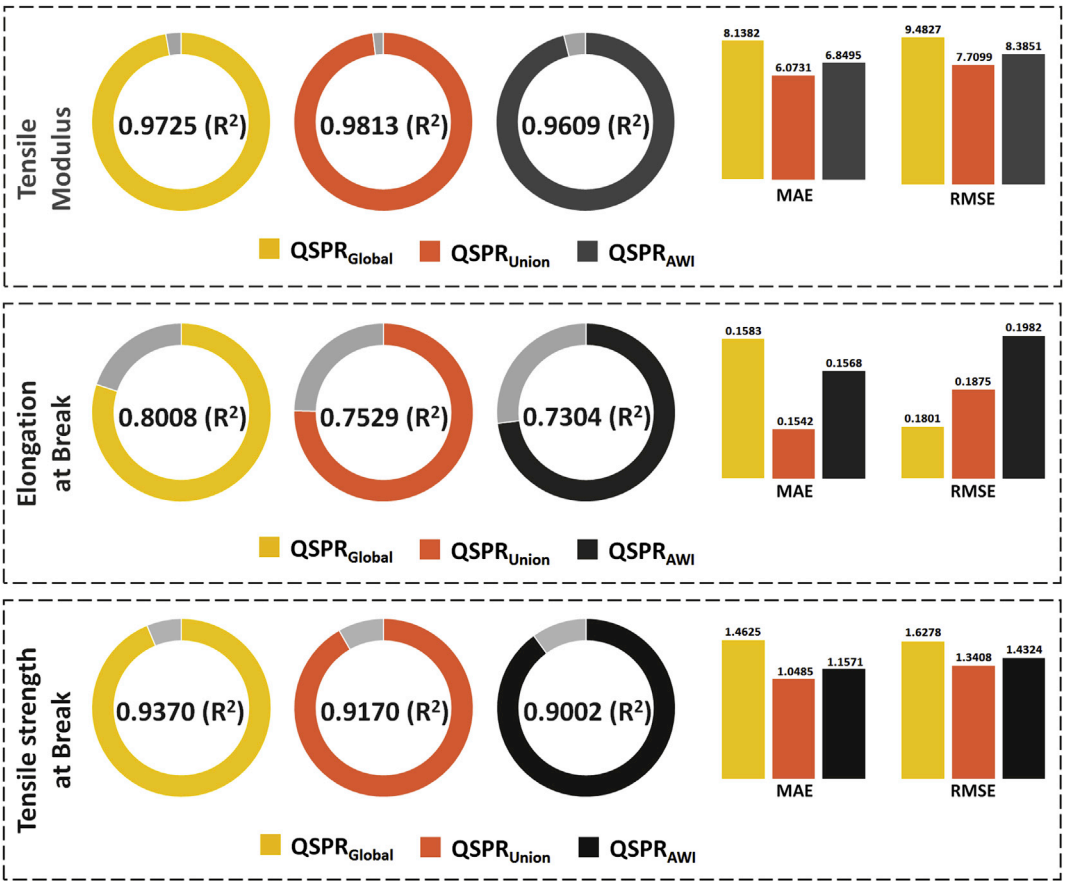

Fig. 4. Comparative results in terms of R^2 and errors (MAE and RMSE) for the $QSPR_{Global}$, $QSPR_{Union}$, and $QSPR_{AWI}$ models for external validation set of tensile modulus, elongation at break and tensile strength at break.

computational modeling of synthetic polymers is related to the polydispersity that characterizes these macromolecular structures. The QSPR models proposed in the literature for predicting material properties avoid this issue, oversimplifying the computational representation of macromolecules as structural repetitive units (SRUs).

Our contributions with this work were to assess the effect of this simplistic vision of the polymer complexities and to propose new ideas for achieving other characterizations of polymers that capture the polydispersity phenomenon, at least partially. In particular, we focused on exploring a key hypothesis: *QSPR models inferred from structural information corresponding to several polymeric chains of different characteristic weights of the materials should yield to more accurate estimations than QSPR models generated from SRU-based representations or other simplified representations*.

A study for estimating three tensile properties of polymers was presented to evaluate the proposed hypothesis. Different computational representations were evaluated in combination with several machine learning techniques. These methods were used for feature selection with the aims of selecting the most relevant MD related to the target property and for inferring the regression methods associated with the QSPR models.

It is clear from our results that the oversimplification of polymer representations is, in general, an unadvisable practice for QSPR modeling, at least under the scope of this case study. Regarding alternative ideas for capturing polydispersity, we contribute with a database of representations based on the calculation of the molecular descriptors for three polymeric chain lengths for each material (SRU, Mn, and Mw), whose models achieve high performances. In particular, it is clear that the QSPR models obtained from databases that included different weight instances of polymers achieve better generalizability skills. As further research, we plan to extend our in-house databases to improve the applicability domain of QSPR models and evaluate the proposed representation strategy for other mechanical polymer properties.

Acknowledgements

This work was supported by Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, [Grant N° PIP 112-2012-0100471]; the Universidad Nacional del Sur (UNS), Bahía Blanca, Argentina, [Grants N° PGI 24/N042 and PGI 24/ZM17]; and the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund [Grant N° TIN2015-64776-C3-2-R DIFERENTIAL @UPO: Massive data management, filtering and exploratory analysis].

We thank Secretaría General de Grado y Educación Continua from the UNS for partially supporting Dr. Barranco's visit to the Instituto de Ciencias e Ingeniería de la Computación (ICIC), Bahía Blanca, Argentina, in 2016, and to the Asociación Universitaria Iberoamericana de Postgrado (AUIP), Salamanca; Spain, for partially supporting Dr. Ponzoni's visit to the Pablo de Olavide University, Sevilla, Spain, in 2017.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2019.06.006.

References

[1] J.B. Mitchell, Machine learning methods in chemoinformatics, Wiley Interdiscip. Rev.: Comput. Mol. Sci. 4 (5) (2014) 468–481.
[2] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. 2 (2016) 16028.

[3] Y. Liu, T. Zhao, W. Ju, S. Shi, Materials discovery and design using machine learning, J. Materiom. 3 (3) (2017) 159–177.
[4] O.R. Bingol, B. Schiefelbein, R.J. Grandin, S.D. Holland, A. Krishnamurthy, An integrated framework for solid modeling and structural analysis of layered composites with defects, Comput. Aided Des. 106 (2019) 1–12.
[5] M. Enciso, N. Meftahi, M.L. Walker, B.J. Smith, BioPPSy: an open-source platform for QSAR/QSPR analysis, PLoS One 11 (11) (2016) e0166298.
[6] 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 (11-12) (2009) 1509–1523.
[7] M.J. Martínez, I. Ponzoni, M.F. Díaz, G.E. Vazquez, A.J. Soto, Visual analytics in cheminformatics: user-supervised descriptor selection for QSAR methods, J. Cheminf. 7 (1) (2015) 39.
[8] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: critical role of the descriptor, Phys. Rev. Lett. 114 (10) (2015) 105503.
[9] A.L. Teixeira, J.P. Leal, A.O. Falcao, Random forests for feature selection in QSPR Models-an application for predicting standard enthalpy of formation of hydrocarbons, J. Cheminf. 5 (1) (2013) 9.
[10] D.W. Van Krevelen, K. Te Nijenhuis, Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions, Elsevier, Oxford, 2009.
[11] T. Le, V.C. Epa, F.R. Burden, D.A. Winkler, Quantitative structure–property relationship modeling of diverse materials properties, Chem. Rev. 112 (5) (2012) 2889–2919.
[12] K. Wu, N. Sukumar, N. Lanzillo, C. Wang, R. Ramprasad, R. Ma, A. Baldwin, G. Sotzing, C. Breneman, Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: toward optimized dielectric polymeric materials, J. Polym. Sci. B Polym. Phys. 54 (20) (2016) 2082–2091.
[13] F. Jabeen, M. Chen, B. Rasulev, M. Ossowski, P. Boudjouk, Refractive indices of diverse data set of polymers: a computational QSPR based study, Comput. Mater. Sci. 137 (2017) 215–224.
[14] D.J. Audus, J.J. de Pablo, Polymer informatics: opportunities and challenges, ACS Macro Lett. 6 (2017) 1078–1082, 2017.
[15] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, npj Comput. Mater. 3 (1) (2017) 54.
[16] C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, Polymer genome: a data-powered polymer informatics platform for property predictions, J. Phys. Chem. C 122 (31) (2018) 17575–17585.
[17] V. Venkatraman, B.K. Alsberg, Designing high-refractive index polymers using materials informatics, Polymers 10 (1) (2018).
[18] N. Adams, Polymer informatics, in: Polymer Libraries, Springer, Berlin, Heidelberg, 2010, pp. 107–149.
[19] F. Cravero, S. Schustik, M.J. Martínez, M.F. Díaz, I. Ponzoni, FS4RVDD: a feature selection algorithm for random variables with discrete distribution, in: J. Medina, M. Ojeda-Aciego, J. Verdegay, I. Perfilieva, B. Bouchon-Meunier, R. Yager (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems. Applications. IPMU 2018. Communications in Computer and Information Science, vol 855, Springer, Cham, Switzerland, 2018.
[20] D. Palomba, G.E. Vazquez, M.F. Díaz, Prediction of elongation at break for linear polymers, Chemometr. Intell. Lab. Syst. 139 (2014) 121–131.
[21] F. Cravero, S. Schustik, M.J. Martínez, C.D. Barranco, M.F. Díaz, I. Ponzoni, Feature selection and polydispersity characterization for QSPR modelling: predicting a tensile property, in: F. Fdez-Riverola, M. Mohamad, M. Rocha, J. De Paz, P. González (Eds.), Practical Applications of Computational Biology and Bioinformatics, 12th International Conference. PACBB2018 2018. Advances in Intelligent Systems and Computing, vol 803, Springer, Cham, 2019.
[22] F. Cravero, M.J. Martínez, G.E. Vazquez, M.F. Díaz, I. Ponzoni, Feature learning applied to the estimation of tensile strength at break in polymeric material design, J. Integr. Bioinf. 13 (2) (2016) 15–29.
[23] M. Froimowitz, HyperChem: a software package for computational chemistry and molecular modeling, Biotechniques 14 (6) (1993) 1010.
[24] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J. Comput. Chem. 32 (7) (2011) 1466–1474.
[25] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36.
[26] D. Weininger, A. Weininger, J.L. Weininger, SMILES. 2. Algorithm for generation of unique SMILES notation, J. Chem. Inf. Comput. Sci. 29 (2) (1989) 97–101.
[27] D. Palomba, Predicción de propiedades de sustancias y materiales de interés en la industria química a través del desarrollo de métodos computacionales, PhD thesis, UNS, Bahía Blanca, Argentina, 2014.
[28] M. Ward, J. Sweeney, An Introduction to the Mechanical Properties of Solid Polymers, second ed., John Wiley & Sons Ltd, England, 2004.
[29] R. Guha, Chemical informatics functionality in R, J. Stat. Softw. 18 (5) (2007) 1–16.
[30] 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.
[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, ACM SIGKDD Explor. Newslett. 11 (1) (2009) 10–18.