Journal of Molecular Graphics and Modelling 38 (2012) 137–147

Contents lists available at SciVerse ScienceDirect

Journal of Molecular Graphics and Modelling

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

Novel descriptors from main and side chains of high-molecular-weight polymers applied to prediction of glass transition temperatures

Damián Palomba[a,b], Gustavo Esteban Vazquez[a], Mónica Fátima Díaz[a,b,*]

[a] Laboratory for Research and Development in Scientific Computing (LIDeCC), DCIC, UNS, Avenida Alem 1253, 8000 Bahía Blanca, Argentina
[b] Planta Piloto de Ingeniería Química (PLAPIQUI) CONICET-UNS, Camino "La Carrindanga", Km. 7, Casilla de Correo 717, 8000 Bahía Blanca, Argentina

ARTICLE INFO

Article history:
Accepted 26 April 2012
Available online 8 May 2012

Keywords:
Structure–property relations
Glass transition temperature
Molecular modeling

ABSTRACT

New descriptors of main and side chains for polymers with high molecular weight are presented in order to predict the glass-transition temperature (T_g) by means of T_g/M ratio. They were obtained by molecular modeling for the middle unit in a series of three repeating units (trimer). Taken together with other classic descriptors calculated for the entire trimeric structure, the ones that correlated better with the property were selected by using a variable selection method. Only three descriptors were chosen: main chain surface area (SA_{MC}), side chain mass (M_{SC}) and number of rotatable bonds (RBN), where the first two descriptors belong to the set of the new ones proposed. By means of a multi-layer perceptron (MLP) neural network a good prediction model (R^2 = 0.953 and RMS = 0.25 K mol/g) was achieved and internally (R^2 = 0.964 and RMS = 0.41 K mol/g) and externally (R^2 = 0.933 and RMS = 0.47 K mol/g) validated. The dataset included 88 polymers. The selected descriptors and the quality of the obtained model demonstrate the advantages of capturing through computational molecular modeling the structural characteristics of the polymers' main and side chains in the prediction of T_g/M.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

The development of new materials with most wanted properties holds great interest for the polymer industry. The ability to predict these properties by in silico methods (i.e., by computer algorithms) is a useful task because the experimental measurement involves the material's synthesis and processing; these activities invariably result in a very time-consuming process and increased costs. In this sense, the glass transition temperature (T_g) is one of the properties of amorphous polymers and composites widely modeled. T_g indicates the temperature below which the material becomes rigid and brittle due to the loss of molecular mobility. This transition is one of the most important characteristics of the material concerning the mechanical and physical properties. The mechanical properties undergo profound changes in the temperature range where this transition occurs [1], thus conditioning the manufacture process and the material's employment.

In this context, models of quantitative structure–property relationship (QSPR) allow to establish relationships between the polymers' structural characteristics and a given physicochemical property, such as T_g. There are two types of QSPR methods currently recognized [1,2] in the prediction of T_g: empirical methods (or the ones that are based on group-counting descriptors) [3–7] and theoretical estimates that use molecular descriptors [1,2,8–23], although sometimes the boundary between these types is not entirely clear [24]. Empirical methods correlate the target property with other chemical and physical properties of polymers, e.g., the theory of additive-group properties [6]. The limitation of these methods is that they are applicable to polymers containing chemical groups previously investigated; however, when combined with molecular modeling, Koehler and Hopfinger [24] were able to estimate the unknown parameters theoretically. On the other hand, theoretical estimations generally employ molecular descriptors based on the structure of the monomer [10,14,18,21,23,25] and/or the repeating unit of the polymer [1,2,9,11–17,20,22]. Some results of the work done on a common dataset are described below as examples of theoretical estimations. Katritzky et al. [1] apply the CODESSA method to predict T_g/M values for 88 linear homopolymers. Five descriptors calculated for the repeating unit were obtained and a QSPR model was generated. García-Domenech and de Julian-Ortiz [25] used 84 polymers of the same database to predict the T_g/M values; they developed a model of 10 parameters generated from graph theoretical indices that were based on the monomers. Cao and Lin [15] designed five descriptors from

* Corresponding author at: Planta Piloto de Ingeniería Química (PLAPIQUI) CONICET-UNS, Camino "La Carrindanga", Km. 7, Casilla de Correo 717, 8000 Bahía Blanca, Argentina. Tel.: +54 291 4861700x255; fax: +54 291 4861600.
E-mail address: mdiaz@plapiqui.edu.ar (M.F. Díaz).

1093-3263/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmgm.2012.04.006

the repeating unit to express the chain stiffness and intermolecular forces of polymers and they correlated them with T_g values. Afantitis et al. [16] achieved an improved correlation coefficient for the same polymers and descriptors by using a radial basis function (RBF) in an artificial neural network (ANN).

In this decade QSPR methods with neural networks have been in vogue. This approach has yielded better results than those achieved with linear methods, such as multilinear regression (MLR) [15,16,22,26,27]. Sumpter and Noid [9] developed an ANN model based on topological indices for a dataset of 320 compounds. On the same dataset they applied the PropNet computational technique to predict T_g, among other properties [13]. Mattioni and Jurs [14] generated two models based on a committee of neural networks with 10 and 11 numerical descriptors based on the monomers and repeating units, respectively. They predicted T_g for two sets of polymers. Duce et al. [17] used a recursive neural network with a hierarchical set of labeled vertexes connected by edges that belong to subclasses of graphs to predict the T_g of (meth)acrylic polymers. Yu et al. [21] utilized a back-propagation artificial neural network (BPANN) for modeling the T_g values of three vinyl polymers kinds with 4 descriptors obtained from the polymer's monomer. Ning [22] employed a BPANN with three descriptors that reflect the chain stiffness derived from the structure of the repeating unit to predict the T_g values of 52 fluorine-containing polybenzoxazoles. Liu and Cao [23] used a BPANN with 4 quantum chemical descriptors obtained from the polymers' monomers, to correlate with the T_g values of 113 polyacrylates and polystyrenes.

The hypothesis of this paper is to propose a new set of descriptors that can establish a better correlation with the target T_g resulting in a predictive model of higher quality and interpretability than the existent ones. While other authors have used different strategies related to the side chain [15,16,18,28] and/or the main chain [22,24] of the polymers, none of them has employed the QSPR approach like our proposal herein. In this work we propose a new perspective in the prediction of the T_g of high molecular weight polymers by directly estimating properties of both fragments (main and side chains) of the repeating unit of a trimeric structure, also taking into account the three dimensional structure that results from interactions with other repeating units. As a result 26 new descriptors were formulated based on physical, chemical, geometrical and electronic features of polymers' main and side chains. In order to supplement the information of these new descriptors, also 635 classic descriptors were calculated for the entire trimer.

A method of variable selection was applied to all descriptors (new and classic ones), with a view to getting some of the new descriptors selected in accordance with the hypothesis. Then, a neural network was developed using those selected descriptors as input in order to obtain the prediction model. Finally a detailed structural explanation of the model descriptors and its relation to the property studied (target) was presented. The methodology is summarized in Fig. 1.

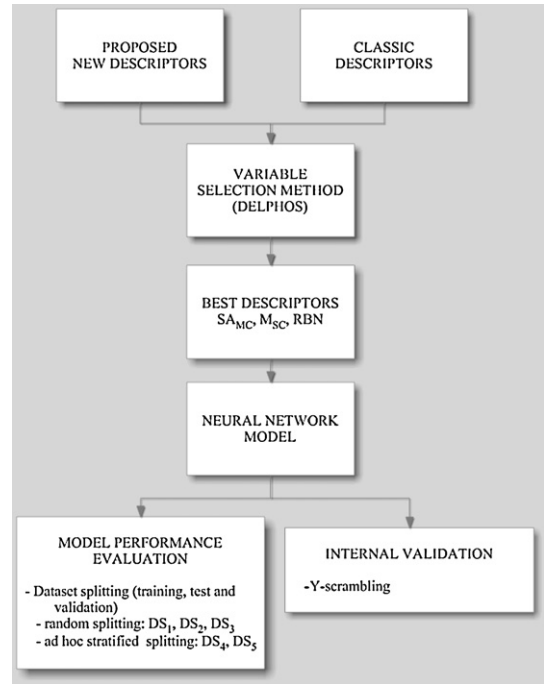

Fig. 1. Scheme of methodology.

2. Experimental

In this paper a QSPR modeling technique for predicting the T_g/M of high molecular weight polymers was applied. The present section is divided in accordance with the usual generation process of a QSPR model: (2.1) Dataset, (2.2) Structure entry and optimization, (2.3) Molecular descriptors generation and (2.4) Model development.

2.1. Dataset

It was found that an alternative way to predict T_g is to obtain the T_g/M ratio (M is the molecular weight of the repeating unit) that results more appropriate for certain study cases [1,25]. According to this strategy, the descriptors are independent of polymer molecular weight [25,29].

In this work the T_g values for uncrosslinked homopolymers in the most common atactic forms were taken from a published compilation [8,25] and were converted to T_g/M (Table 1). This dataset consists of different polymer families in order to ensure a high level of structural heterogeneity (polyethylenes, polyacrylates, polymethacrylates, polystyrenes, polyethers, polyoxides and molecules containing functional groups, such as halides, cyanide, carboxylates, acetates, amides, ethers, and alcohols). There are also molecules with hydrocarbon side chains and aliphatic and aromatic rings. The values of T_g/M range from 1.07 to 8.14 K mol/g.

2.2. Structure entry and optimization

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 [30] (Hypercube, Inc.) (Fig. 2). The molecules were optimized with 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 due to intramolecular forces. At first, the structures were pre-optimized with the Force Field Molecular Mechanics (MM+) procedure; then, the resulting geometries were further refined by means of the Semi-Empirical Molecular Orbitals

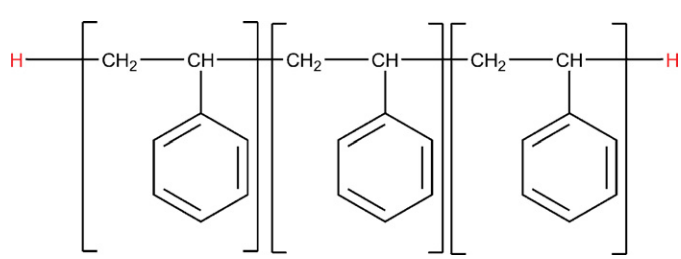

Fig. 2. A sample of trimeric molecular model for polystyrene.

Table 1
Dataset including observed (exp.) glass transition temperature ratio (T_g/M) and their corresponding predicted (calc.) values for random and ad hoc dataset splittings.

N	Name of polymers	T_g/M (K mol/g)					
		(Exp.)	Random data splitting (calc.)			Ad hoc data splitting (calc.)	
			DS1[a] 50–23–27[b]	DS2[a] 60–20–20[b]	DS3[a] 50–25–25[b]	DS4[a] 60–15–25[b]	DS5[a] 60–15–25[b]
#1	Poly(ethylene)	6.96	7.13[v]	7.13[v]	7.12[v]	6.97[c]	6.83[c]
#2	Poly(ethylethylene)	4.07	4.4[c]	4.1[c]	3.83[t]	4.14[t]	5.17[v]
#3	Poly(butylethylene)	2.62	2.96[v]	3.02[v]	3.12[v]	3.15[c]	2.56[c]
#4	Poly(cyclopentylethylene)	3.63	3.55[t]	3.38[c]	3.44[t]	3.32[t]	3.99[v]
#5	Poly(cyclohexylethylene)	3.3	3.38[c]	3.21[c]	3.32[c]	3.1[c]	3.49[c]
#6	Poly(acrylic acid)	5.26	3.9[t]	4.22[t]	3.67[t]	3.75[c]	4.53[c]
#7	Poly(methyl acrylate)	3.27	3.28[v]	3.25[v]	3.31[v]	3.29[c]	3.36[c]
#8	Poly(ethyl acrylate)	2.51	2.76[v]	2.75[t]	2.84[v]	2.87[c]	2.38[c]
#9	Poly(sec-butyl acrylate)	1.98	2.19[c]	2.17[c]	2.08[c]	2.28[v]	1.78[t]
#10	Poly(vinyl alcohol)	8.14	7.13[t]	7.06[t]	7.13[t]	7.27[v]	6.04[t]
#11	Poly(vinyl chloride)	5.57	5.56[c]	5.69[c]	5.56[c]	5.57[t]	4.96[v]
#12	Poly(acrylonitrile)	7.13	7.13[c]	7.04[c]	7.12[c]	7.22[c]	5.48[c]
#13	Poly(vinyl acetate)	3.5	3.28[c]	3.24[c]	3.31[c]	3.28[c]	3.35[c]
#14	Poly(styrene)	3.59	3.41[t]	3.25[t]	3.32[t]	3.16[c]	3.76[c]
#15	Poly(2-chlorostyrene)	2.84	2.93[t]	2.76[c]	2.76[t]	2.58[v]	2.64[v]
#16	Poly(3-chlorostyrene)	2.63	2.93[c]	2.75[c]	2.76[c]	2.58[c]	2.64[c]
#17	Poly(4-chlorostyrene)	2.82	2.93[v]	2.76[v]	2.77[v]	2.58[t]	2.63[v]
#18	Poly(2-methylstyrene)	3.47	3.22[c]	3.05[c]	3.12[c]	2.92[v]	3.33[v]
#19	Poly(3-methylstyrene)	3.17	3.21[c]	3.04[c]	3.1[c]	2.91[v]	3.34[t]
#20	Poly(4-methylstyrene)	3.17	3.2[c]	3.04[c]	3.1[c]	2.91[t]	3.34[v]
#21	Poly(4-fluorostyrene)	3.11	3.15[v]	2.98[v]	3.03[v]	2.84[c]	3.22[c]
#22	Poly(propylene)	5.55	7.13[v]	5.85[t]	7.13[v]	5.54[c]	6.14[c]
#23	Poly(1-pentene)	3.14	3.55[v]	3.9[v]	3.55[v]	3.62[c]	3.83[c]
#24	Poly(ethoxyethylene)	3.53	3.47[c]	3.76[c]	3.47[c]	3.53[v]	3.48[v]
#25	Poly(tert-butyl acrylate)	2.46	2.44[v]	2.39[v]	2.35[v]	2.42[c]	2.16[c]
#26	Poly(n-butyl acrylate)	1.71	1.99[t]	1.98[t]	1.87[t]	2.15[v]	1.6[v]
#27	Poly(vinyl hexyl ether)	1.63	1.8[c]	1.8[c]	1.65[c]	2.01[t]	1.51[v]
#28	Poly(1,1-dimethylethylene)	3.55	3.87[c]	3.52[c]	4[c]	3.58[c]	4.35[c]
#29	Poly(1,1-dichloroethylene)	2.64	3.05[v]	3.81[v]	3.75[v]	3.63[c]	2.85[c]
#30	Poly(1,1-difluoroethylene)	3.64	3.77[c]	3.64[c]	3.95[c]	3.65[c]	3.98[c]
#31	Poly(a-methylstyrene)	3.47	4.67[v]	3.18[v]	3.36[v]	3.06[c]	2.62[c]
#32	Poly(methyl methylacrylate)	3.78	3.23[c]	3.15[c]	3.37[c]	3.19[c]	3.9[c]
#33	Poly(ethyl methylacrylate)	2.84	2.72[c]	2.7[c]	2.91[c]	2.78[v]	3.09[v]
#34	Poly(isopropyl methylacrylate)	2.55	2.56[v]	2.51[t]	2.66[t]	2.55[c]	2.86[c]
#35	Poly(ethyl chloroacrylate)	2.73	2.5[v]	2.45[t]	2.59[v]	2.47[c]	2.83[c]
#36	Poly(2-chloroethyl methylacrylate)	2.47	2.35[c]	2.26[c]	2.31[c]	2.26[c]	2.52[c]
#37	Poly(tert-butyl methylacrylate)	2.68	2.41[c]	2.34[c]	2.42[c]	2.34[c]	2.63[c]
#38	Poly(phenyl methylacrylate)	2.43	2.22[c]	2.12[c]	2.1[c]	2.08[c]	2.27[c]
#39	Poly(chlorotrifluoroethylene)	3.22	3.07[t]	3.32[c]	3.85[t]	3.57[c]	3.12[c]
#40	Poly(oxymethylene)	7.27	7.13[t]	6.89[t]	7.13[t]	7.27[c]	6.95[c]
#41	Poly(oxyethylene)	4.68	4.23[c]	4.58[c]	4.73[c]	4.17[t]	5.1[v]
#42	Poly(oxytrimethylene)	3.36	3.38[c]	3.35[c]	3.46[c]	3.54[v]	3.47[v]
#43	Poly(oxytetramethylene)	2.64	2.69[t]	2.74[c]	2.6[t]	2.97[c]	2.56[c]
#44	Poly(ethylene terephthalate)	1.8	1.51[t]	1.47[t]	1.88[t]	1.54[c]	1.48[c]
#45	Poly(vinyl n-octyl ether)	1.24	1.41[v]	1.39[v]	1.28[v]	1.51[c]	1.46[c]
#46	Poly(vinyl n-decyl ether)	1.07	1.17[c]	1.18[c]	1.17[c]	1.16[c]	1.46[c]
#47	Poly(oxyoctamethylene)	1.59	1.32[c]	1.32[c]	1.87[c]	1.43[c]	1.54[c]
#48	Poly(oxyhexamethylene)	2.04	1.79[c]	1.8[c]	1.93[c]	2.06[v]	1.76[v]
#49	Poly(vinyl n-pentyl ether)	1.82	2.09[c]	2.1[c]	2.02[c]	2.32[v]	1.6[t]
#50	Poly(vinyl 2-ethylhexyl ether)	1.33	1.52[c]	1.49[c]	1.35[c]	1.61[c]	1.48[c]
#51	Poly(n-octyl acrylate)	1.13	1.24[c]	1.23[c]	1.19[c]	1.23[c]	1.46[c]
#52	Poly(n-octyl methylacrylate)	1.28	1.23[c]	1.22[c]	1.2[c]	1.19[c]	1.46[c]
#53	Poly(n-heptyl acrylate)	1.25	1.36[v]	1.34[v]	1.25[v]	1.41[t]	1.47[v]
#54	Poly(n-nonyl acrylate)	1.09	1.15[t]	1.16[c]	1.16[t]	1.08[c]	1.46[c]
#55	Poly(n-hexyl acrylate)	1.38	1.52[t]	1.5[c]	1.36[t]	1.62[c]	1.48[c]
#56	Poly(1-heptene)	2.24	2.5[t]	2.54[t]	2.61[t]	2.74[v]	1.91[t]
#57	Poly(vinyl n-butyl ether)	2.21	2.45[c]	2.48[c]	2.51[c]	2.68[v]	1.83[t]
#58	Poly(n-propyl acrylate)	2.01	2.33[c]	2.33[c]	2.32[c]	2.49[t]	1.84[v]
#59	Poly(vinylisobutyl ether)	2.56	2.74[c]	2.74[c]	2.81[c]	2.85[c]	2.3[v]
#60	Poly(vinyl sec-butyl ether)	2.53	2.75[t]	2.75[c]	2.83[t]	2.86[c]	2.35[c]
#61	Poly(pentafluoroethyl ethylene)	2.15	2.13[c]	2.66[c]	2.64[c]	2.47[c]	2.33[c]
#62	Poly(2,3,3,3-tetrafluoropropylene)	2.76	2.57[v]	3.48[v]	3.56[v]	3.27[c]	2.51[c]
#63	Poly(3,3-dimethylbutyl methacrylate)	1.87	1.78[v]	1.71[v]	1.6[v]	1.76[t]	1.71[v]
#64	Poly(n-butyl acrylamide)	2.51	2.22[t]	2.19[t]	2.14[t]	2.31[c]	1.83[c]
#65	Poly(vinyl trifluoroacetate)	2.28	2.56[v]	2.45[v]	2.36[v]	2.37[c]	2.47[c]
#66	Poly(3-methyl-1-butene)	4.61	4.42[c]	4.55[c]	3.74[c]	3.84[v]	4.74[t]
#67	Poly(n-butyl a-chloroacrylate)	2.04	1.81[c]	1.75[c]	1.65[c]	1.83[c]	1.7[c]
#68	Poly(sec-butyl methacrylate)	2.32	2.17[c]	2.13[c]	2.15[c]	2.21[v]	2.15[t]
#69	Poly(heptafluoropropyl ethylene)	1.69	2.07[v]	1.89[v]	1.67[v]	1.72[v]	1.47[t]
#70	Poly(3-pentyl acrylate)	1.81	1.88[t]	1.85[c]	1.7[t]	1.97[t]	1.58[v]
#71	Poly(5-methyl-1-hexene)	2.64	2.79[c]	2.79[c]	2.89[c]	2.91[t]	2.43[v]
#72	Poly(oxy-2,2-dichloromethyltrimethylene)	2.09	2.82[c]	2.72[c]	2.55[c]	2.66[c]	2.16[c]

Table 1 (*Continued*)

N	Name of polymers	T_g/M (K mol/g)					
		(Exp.)	Random data splitting (calc.)			Ad hoc data splitting (calc.)	
			DS1[a] 50–23–27[b]	DS2[a] 60–20–20[b]	DS3[a] 50–25–25[b]	DS4[a] 60–15–25[b]	DS5[a] 60–15–25[b]
#73	Poly(n-hexyl methacrylate)	1.58	1.51[c]	1.48[c]	1.38[c]	1.57[v]	1.51[t]
#74	Poly(vinyl isopropyl ether)	3.14	3.28[t]	3.24[c]	3.31[t]	3.28[c]	3.34[c]
#75	Poly[p-(n-butyl)styrene]	1.74	1.95[v]	1.87[v]	1.71[v]	1.89[c]	1.73[c]
#76	Poly(n-butyl methacrylate)	2.06	1.97[t]	1.95[t]	1.94[t]	2.09[c]	1.81[c]
#77	poly(2-methoxyethyl methacrylate)	2.03	1.96[c]	1.93[c]	1.91[c]	2.06[v]	1.8[t]
#78	Poly(3,3,3-trifluoropropylene)	3.13	2.75[c]	3.63[c]	3.6[c]	3.48[t]	3.53[v]
#79	Poly(4-methyl-1-pentene)	3.64	3.34[c]	3.34[c]	3.39[c]	3.36[c]	3.61[c]
#80	Poly(vinyl chloroacetate)	2.53	2.81[v]	2.72[v]	2.75[v]	2.68[v]	2.83[t]
#81	Poly(n-propyl methacrylate)	2.39	2.3[c]	2.29[c]	2.4[c]	2.41[v]	2.28[v]
#82	poly(3-cyclopentyl-1-propene)	3.03	2.99[v]	2.91[t]	3.03[v]	2.91[c]	3.27[c]
#83	Poly(3-phenyl-1-propene)	2.82	2.86[c]	2.77[c]	2.82[c]	2.74[v]	2.96[t]
#84	Poly(n-propyl a-chloroacrylate)	2.32	2.11[t]	2.05[t]	2.03[t]	2.11[c]	2.06[c]
#85	Poly(sec-butyl a-chloroacrylate)	2.14	2[t]	1.92[t]	1.84[t]	1.94[c]	1.97[c]
#86	Poly(3-cyclohexyl-1-propene)	2.81	2.82[v]	2.72[t]	2.79[v]	2.68[v]	3.03[c]
#87	Poly(vinyl acetal)	3.11	3.69[c]	3.48[c]	3.46[c]	3.45[c]	2.72[c]
#88	Poly(vinyl formal)	3.78	3.62[c]	3.44[c]	3.19[c]	3.43[c]	4.07[c]

N = number of molecule. c is the calibration/training set; t is the test set; v is the validation set.
[a] The DSi correspond to dataset splittings *i*, with *i* = 1, . . ., 5.
[b] The values correspond to the percentage of dataset members that belong to the calibration, test and validation sets, respectively.

Method AM 1 (Austin Model 1) by using Polak–Ribiere's algorithm and a gradient norm limit of 0.01 $kcal^{-1}\ mol^{-1}$.

2.3. Generation of the molecular descriptors

2.3.1. Calculation of the new descriptors

The fragments (main and side chains) considered in this work belong to the middle repeating unit. In this work the main chain has been defined as the succession of all atoms (also including the hydrogens attached to them) that are in the backbone of the trimer middle repeating unit. The remaining atoms in this middle repeating unit are considered as the side chain, thus avoiding ambiguities.

Some cases of main and side chains are shown in Fig. 3. For example, the main chain in poly(ethylene) (PE) is defined by two carbon atoms and four hydrogen atoms, while in poly(styrene) (PS) it is defined by two carbon atoms and three hydrogen atoms. Likewise, the side chain for PE is null (properties for this fragment are considered equal to zero), while for the PS it is the phenyl group. The particular cases in Fig. 3 are poly(ethylene terephthalate), poly(vinyl formal) and poly(vinyl acetal). For the poly(ethylene terephthalate) the main chain is defined by two carbon atoms, four hydrogen atoms, a phenyl group, two carbon atoms and two oxygen atoms. The side chain is composed of two oxygen atoms that belong to the carbonyls. For poly(vinyl formal) the whole middle repeating unit is considered as the main chain (the side chain is null); for the poly(vinyl acetal) the main chain is defined in the same way, but the side chain is the methyl group hanging from the cycle.

Once the molecules had been drawn and optimized, the following 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: the calculation of Van der Waals surface areas is carried out by an approximate method [31,32]. The calculation is fast, and generally accurate to within 10% for a given set of atomic radii. The calculation is based on atomic radii. Hydrogens attached to carbon atoms are not considered explicitly, but are implicitly included with their carbon atom.
- Van der Waals volume: this calculation employs the grid method described by Bodor et al. [33,34].
- Log *P* (logarithm octanol–water partition coefficient): log *P* calculation is carried out by using atomic parameters derived by Ghose et al. [35] and later extended by Ghose and coworkers [36]. It is well known that log *P* is experimentally estimated [37]; for example, with the traditional Shake-Flask method for the entire molecule. By using this fragment contribution method, the log *P* for a fraction of the molecule can be estimated as a measure of the fragment's hydrophobicity.
- Refractivity: this property is estimated by the same method as the one for log *P*. Ghose and Crippen presented atomic contributions to the refractivity in exactly the same manner as to the hydrophobicity [36,38].
- Polarizability: it is estimated from an additivity scheme presented by Miller [39]. Different increments are associated with diverse atom types.
- Mass: mass of the polymer fragment considered.
- Number of atoms: number of atoms in the polymer fragment considered.

In other words, these specific properties were estimated for fragments of polymers thus obtaining 14 descriptors, 7 for the main chain and 7 for side chain. Then, the same properties 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 side chain. In Table 2,

Table 2
Nomenclature of new descriptors.

New descriptors	Nomenclature		Units
	Unnormalized	Normalized	
Main chain surface area	SA_{MC}	nSA_{MC}	$Å^2$
Main chain volume	V_{MC}	nV_{MC}	$Å^3$
Main chain log *P*	$Log\,P_{MC}$	$nLog\,P_{MC}$	–
Main chain refractivity	R_{MC}	nR_{MC}	$Å^3$
Main chain polarizability	P_{MC}	nP_{MC}	$Å^3$
Main chain mass	M_{MC}	nM_{MC}	Da
Main chain atoms number	N_{MC}	–	–
Side chain surface area	SA_{SC}	nSA_{SC}	$Å^2$
Side chain volume	V_{SC}	nV_{SC}	$Å^3$
Side chain log *P*	$Log\,P_{SC}$	$nLog\,P_{SC}$	–
Side chain refractivity	R_{SC}	nR_{SC}	$Å^3$
Side chain polarizability	P_{SC}	nP_{SC}	$Å^3$
Side chain mass	M_{SC}	nM_{SC}	Da
Side chain atoms number	N_{SC}	–	–

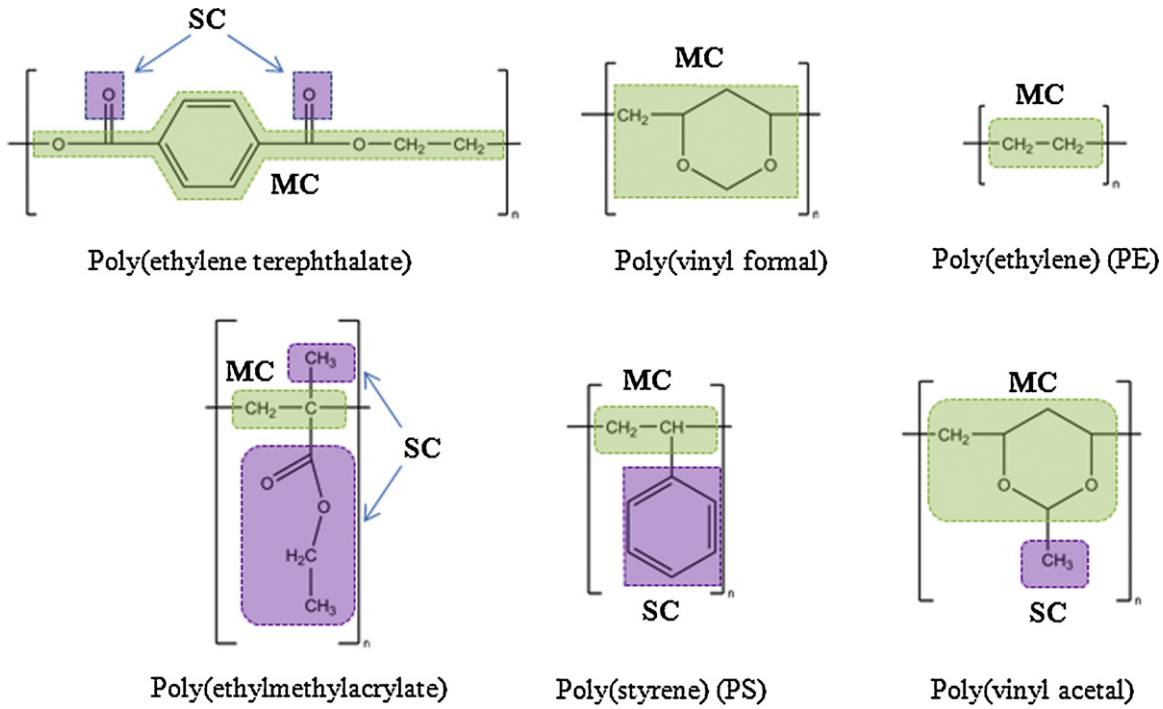

Fig. 3. Identification of the fragments of trimeric design. Examples of main chain (MC) and side chain (SC) fragments that belong to the repeating unit with trimeric structure.

the nomenclature for the 26 descriptors proposed is presented and their values are available as a Supplementary file.

2.3.2. Calculation of the classic descriptors

As a supplement of the new descriptors that were proposed, a set of classic variables were calculated considering the whole trimer by using Dragon 5.5 software [29,40]. Some descriptors were not considered. In addition to all binary descriptors, fingerprints (2D binary and 2D frequency fingerprints) [40,41] were not considered to avoid the introduction of the "missing structure" phenomenon [42]; e.g., the missing structures might be fragments, functional groups, etc. This is due to problems the QSPR models have when some fragments do not exist in the training set, or when they have a very low frequency and thus, the coefficients associated with these sub-structures are not significant statistically [42]. Some descriptors belonging to the Molecular Properties category [40,41] were excluded since they are related to drugs characteristics (e.g., drug like index [41]), which are obviously completely different from polymer features. The 3D descriptors [40] were also avoided in order to obtain simpler models. Finally, constants' descriptors (i.e., all the variables that take the same value for all samples in the dataset) and near constants (i.e., variables that assume the same value, except for one or very few cases) were deleted. The final pool of classic descriptors chosen consisted of 635 variables and their values are available as a Supplementary file.

2.4. Model development

The model building process included as a first stage is a variable selection method. The objective of variable selection is to reduce the set of descriptors (independent variables) when predicting a target (T_g/M) with the aim of improving the prediction performance of the descriptors and providing a better understanding of the underlying process. This technique was applied to the whole set of descriptors proposed in Sections 2.3.1 and 2.3.2. Then, the selected descriptors were used as input for an ANN. Various validation tests were carried out to ensure model correctness. Methodology was summarized in Fig. 1.

2.4.1. Variable selection

Delphos [43], which is a piece of software for linear and nonlinear feature extraction, was employed for the selection of the most representative descriptors. Delphos is based on a wrapper methodology that works as follows: in a first phase a genetic algorithm (GA) is used in order to find the best subsets of descriptors where the fitness function enables different regression techniques to assess these subsets. In a second stage, the best subsets are rigorously evaluated by an ANN. As a result, Delphos provides as output multiple sets of descriptors best correlated with the target property, based on the lowest mean absolute error (MAE) and mean square error (MSE). Finally, the user has the facility to choose among these sets of descriptors; this selection can be based either on the prediction accuracy (minimum prediction error), the physicochemical meaning, the interpretability of selected variables, the number of selected descriptors, among others, or a combination of them.

In this work, all of these criteria were used and the best set of variables selected in correlation with T_g/M consisted of three descriptors. As expected according to the hypothesis, two of them correspond to the new descriptors proposed: SA_{MC} (main chain surface area), M_{SC} (side chain mass). Besides, the classic RBN descriptor (number of rotatable bonds) was also selected. These descriptors' numerical values are shown in Table 3.

2.4.2. Nonlinear modeling with ANNs

The best set of descriptors, which was mentioned in the previous section, was used as input in a multi-layer neural network perceptron (MLP) for the same target (T_g/M) by using STATISTICA 8.0 software [44]. The network architecture was adjusted by trial and error to achieve optimal performance: MLP 3–3–1 (three input layer neurons, three hidden layer neurons and one output layer neuron), with the activation function Tanh (Hyperbolic Tangent) for both the hidden and output layers, the error function SOS (sum

Table 3
Descriptors' numerical values used in this work.

N^a	SA_{MC}	M_{SC}	RBN
#1	40.982	0	3
#2	26.539	29.062	7
#3	27.522	57.115	13
#4	26.163	69.126	7
#5	23.718	83.153	7
#6	28.074	45.018	7
#7	28.067	59.045	10
#8	27.872	73.072	13
#9	28.334	101.125	16
#10	30.58	17.007	4
#11	30.612	35.453	3
#12	28.636	26.018	4
#13	28.2	59.045	10
#14	26.771	77.106	7
#15	26.675	111.551	7
#16	26.77	111.551	7
#17	26.462	111.551	7
#18	26.291	91.133	7
#19	26.784	91.133	7
#20	26.913	91.133	7
#21	26.923	95.096	7
#22	27.492	15.035	4
#23	26.796	43.089	10
#24	29.179	45.061	10
#25	28.039	101.125	13
#26	28.081	101.125	19
#27	29.639	101.169	22
#28	18.246	30.07	4
#29	21.078	70.906	3
#30	21.938	37.997	3
#31	17.425	92.141	7
#32	18.551	74.079	10
#33	18.363	88.106	13
#34	18.44	102.133	13
#35	17.259	108.525	13
#36	18.276	122.551	13
#37	18.339	116.16	13
#38	18.021	136.15	13
#39	3.132	92.448	3
#40	35.546	0	3
#41	56.243	0	6
#42	76.924	0	9
#43	97.64	0	12
#44	154.973	31.999	18
#45	29.643	129.222	28
#46	29.652	157.276	34
#47	180.476	0	24
#48	139.055	0	18
#49	29.461	87.142	19
#50	28.596	129.222	25
#51	27.694	157.233	31
#52	18.328	172.268	31
#53	27.788	143.206	28
#54	27.756	171.26	34
#55	27.669	129.179	25
#56	27.503	71.142	16
#57	29.44	73.115	16
#58	27.878	87.098	16
#59	28.863	73.115	13
#60	28.282	73.115	13
#61	26.36	119.014	7
#62	19.324	88.005	4
#63	18.305	144.214	19
#64	26.974	100.141	16
#65	29.665	113.016	10
#66	25.282	43.089	7
#67	19.624	136.578	19
#68	18.46	116.16	16
#69	26.22	169.022	10
#70	28.386	115.152	19
#71	27.496	71.142	13
#72	40.37	85.941	9
#73	17.956	144.214	25
#74	28.258	59.088	10
#75	26.741	133.213	16
#76	17.923	116.16	19
#77	17.971	118.133	19
#78	27.219	69.006	4
#79	26.292	57.115	10
#80	28.276	93.49	10
#81	18.081	102.133	16
#82	25.216	83.153	10
#83	27.282	91.133	10
#84	19.254	122.551	16
#85	19.174	136.578	16
#86	25.065	97.18	10
#87	82.859	15.035	4
#88	98.629	0	4

[a] Numbers of the dataset molecules (Table 1).

of squares) and the BFGS (Broyden–Fletcher–Goldfarb–Shanno) quasi-Newton training algorithm.

The dataset was randomly split into three separate sections: training, test, and validation sets. The training and test sets were used to adjust the model parameters by using an early stopping scheme for regularization; the predictive accuracy of the model was evaluated by using the external validation set.

Three splittings of the original dataset were generated by using different proportions for training, test and validation sets. The proportion of data splitting (DS) consisted in: 50%–23%–27% (DS1), 60%–20%–20% (DS2) and 50%–25%–25% (DS3). All compounds were assigned to each partition randomly. Other dataset partitions (DS4 and DS5) were formulated by using an ad hoc stratified partitioning scheme. The purpose of the stratified sampling is to guarantee a fair distribution of compounds belonging to all polymer families among the training, test and validation sets. Both DS4 and DS5 consisted in a 60%–15%–25% splitting percentage. In order to minimize the risk of chance correlations, a minimum of 20 polymers were kept in the validation and training datasets, following the relationship "cardinality of the set $\geq 5 \cdot$ number_of_descriptors" [45]. All the DSi splittings (including which polymer was assigned to each set) are shown in Table 1. Thus, 5 ANNs were trained and validated by using each DSi set. The y-scrambling technique was applied in order to avoid the possibility of chance correlation of the descriptors. The results are shown in Figs. 4 and 5.

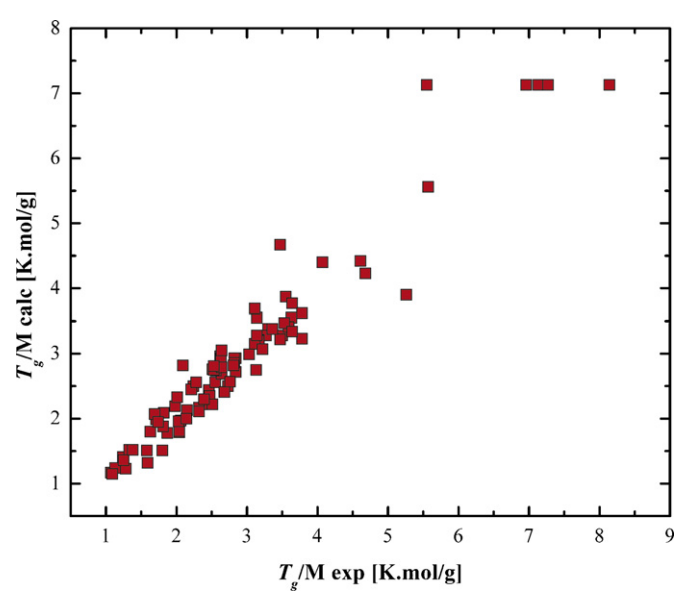

Fig. 4. Calculated vs. experimental values of T_g/M (ANN trained by using DS1).

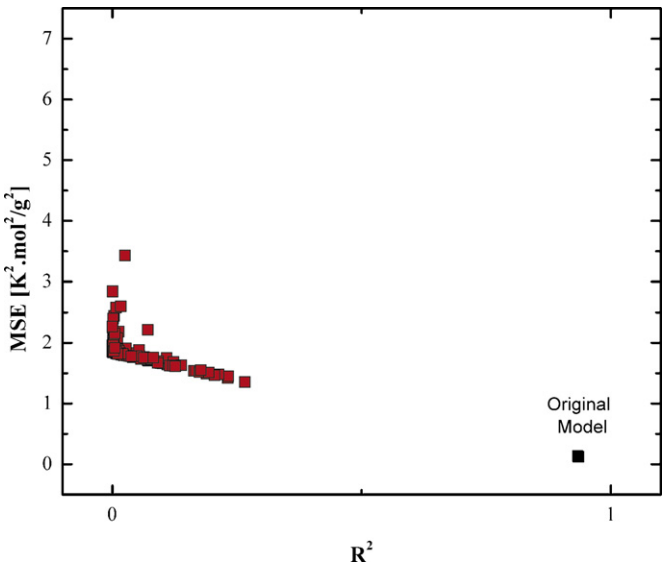

Fig. 5. Y-scrambling. R^2 values from 100 models obtained by randomization of the target values (100 runs).

3. Results and discussion

As has already been stated in the literature [46], chain flexibility, molecular structure and branching are the main factors that affect the polymer's T_g. In order to quantify different properties of polymers, their structures were selectively analyzed: on the one hand, the main chain and on the other hand, the side chain. As a consequence, new descriptors arose from both the main chain and the side chain (Table 2) with a novel approach. The aim was to investigate the relationship between these descriptors and the above mentioned factors. As it has also been cited by other authors [22,23], it is impossible to calculate the descriptors for the entire molecule because all polymer's molecular weights are too high. Thus, a reduced molecular design consisting of a trimer was used to represent each polymer. The advantage of working with a trimer resides 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), as it is influenced by physicochemical, steric and electronic features of adjacent units and also preserves the structural characteristics of the polymer. For example, the middle repeating unit of poly(ethylene terephthalate) trimer (Fig. 3) is the only one that represents the whole molecule since the lateral repeating units are both different: one contains a hydroxyl group and the other one does not. If the descriptors were calculated on this unit with the hydroxyl group, the obtained values would be completely different from those that belong to the middle repeating unit and they would not reflect what really occurs in the original polymer structure. For these reasons, the descriptors were calculated for the middle repeating unit of the trimer, which had previously been optimized. Although a trimer is a very simple representation of a polymer, it is valid to optimize its geometry in order to consider its intramolecular interactions. This molecular optimization does not intend to emulate the 3D conformation of polymer molecule, but to consider intramolecular interactions between atoms of neighbor repeating units in minimum scale. Moreover, the values of the proposed descriptors are affected by these interactions.

In Section 3.1 the performance of the prediction model is presented. In Section 3.2 a physicochemical explanation of the QSPR model is discussed, thus making an effort to enhance its comprehension. Finally in Section 3.3 an evaluation of variable relevance is shown.

3.1. Model performance

Five ANNs, which were calibrated with the training DSi sets, were built through STATISTICA by using the descriptors set reported in Section 2.4.1. The values of observed and predicted T_g/M for the five dataset splits are shown in Table 1. As it can be observed, all models presented a very good performance, according to R^2 (squared correlation coefficient) and other classical statistical parameters (Tables 4 and 5). Although all models resulting from the different splits are remarkably good, model DS1 was chosen due to its lowest mean relative error (MRE). Fig. 4 plots the calculated T_g/M values against the experimental values for this model.

It is important to note that only 3 descriptors were used in the model, following the principle of parsimony (Occam) [47], although the dataset consisted of structurally diverse compounds, thus demonstrating the generalization ability of model's descriptors. Typically, models with greater number of descriptors than the one presented here are proposed in the literature [1,14,48], even in the case studied in [18] where a few compounds in the validation set were employed. Other papers have proposed models that have been generated using few descriptors (two and three), but they are restricted to a particular family of polymers [19,22], which are obviously structurally similar. All these features highlight the good quality of our model. Indeed, only 3 descriptors conform the predictive model exhibiting a similar performance to the ones mentioned above.

By means of the models DS1–DS3, whose data splittings were randomly generated by varying the proportions of the different sets (Table 4), it was shown that the good correlation between the model descriptors and the target property is independent of the number of polymers that constitute each dataset. Furthermore, it was confirmed that the results are still good through DS4 and DS5 ad hoc models (Table 5), even though the sets have an equitable distribution of all polymer families that compose the dataset and the chance factor has been eliminated. Thus, in each of training, test and validation datasets corresponding to DS4 and DS5, structurally different polymeric attributes are captured and the results are not fitted to any family in particular. It is worth mentioning that when working with both randomly selected and ad hoc datasets, no outlier at all was excluded as researches had done in the other reported studies that employed the same dataset [1,16].

As mentioned above, it is advisable to complete the task with a proper validation. To achieve this aim, y-scrambling (internal validation method) was applied, which has scarcely been reported in the literature, but proves to be really useful in order to complement the external validation. The results for our best model (DS1) can be seen in Fig. 5, where all models generated by randomization of target values gave a very poor performance, thus confirming that there was no chance correlation between the model descriptors and the T_g/M values.

3.2. Physicochemical aspects

As indicated in [49], "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". Despite 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 [50].

The aim of this section is to analyze in detail the relationship among descriptors, molecular structure and the target, in order to provide some physicochemical justification of the resulting model. In general, the values obtained from the analyzed fragment descriptors were affected by the presence of adjacent groups, which is

Table 4
Performance of models DS1, DS2 and DS3, from MLP ANN (architecture: 3–3–1). Sets: Training (calibration); the test and validation correspond to random data splitting.

Model	Set	%	R^2	MAE	MSE	RMS	MRE
DS1	Training	50	0.953	0.20	0.06	0.25	7.99
	Test	23	0.964	0.26	0.17	0.41	8.23
	Validation	27	0.933	0.28	0.22	0.47	9.81
DS2	Training	60	0.954	0.20	0.06	0.24	8.00
	Test	20	0.964	0.34	0.19	0.44	9.97
	Validation	20	0.923	0.29	0.17	0.41	11.29
DS3	Training	50	0.941	0.22	0.08	0.28	8.29
	Test	25	0.940	0.30	0.22	0.47	9.12
	Validation	25	0.926	0.28	0.23	0.48	9.04

Table 5
Performance of models DS4 and DS5 from MLP ANN (architecture: 3–3–1). Sets: Training (calibration); the test and validation correspond to ad hoc data splitting.

Model	Set	%	R^2	MAE	MSE	RMS	MRE
DS4	Training	60	0.929	0.26	0.13	0.37	10.01
	Test	15	0.949	0.25	0.09	0.29	10.24
	Validation	25	0.955	0.26	0.13	0.36	9.29
DS5	Training	60	0.921	0.27	0.15	0.39	10.24
	Test	15	0.918	0.36	0.39	0.62	10.81
	Validation	25	0.914	0.27	0.12	0.35	9.26

considered relevant when it is describing such complex molecules. In addition, it was found that there are correlations between the different model descriptors and the target values, which complement each other (see definition and calculations of each descriptor in Section 2.3.1).

3.3. Main chain surface area (SA_{MC})

In case when the polymer structure does not present side groups and/or it has variations on the main chain, an inverse relationship between SA_{MC} and T_g/M is observed: i.e., the larger the area of the main chain, the smaller the T_g/M (Fig. 6). This can be illustrated with the case when the repeating unit has an equal amount of matter, namely molecules that either are isomers or differ in its structure by only a few atoms of hydrogen; some examples are shown in Table 6. When the main chains are more flexible, especially those with free rotation, they occupy more surface area and T_g/M decreases. The same trend is observed for some poly(acrylate) and poly(methyl acrylate) isomers, which possess structural variation on the side chain and do not exhibit disparities in the main chain (Table 6). For example, poly(ethyl acrylate), whose side chain is constituted by only one substituent, and its isomer poly(methyl methylacrylate), which has two substituents on the same main chain, exhibit a lower SA_{MC} and a higher T_g/M. This fact accords with the physical behavior that shows that the latter isomer is stiffer. Although the structure of the main chain remains unchanged, this descriptor's value captures the presence of substituents due to the calculation method (Section 2.3.1). In addition, this descriptor allows to distinguish molecules that only differ structurally in their main chain, such as polyoxides #40, #41, #42, #43, #47 and #48 (Table 1).

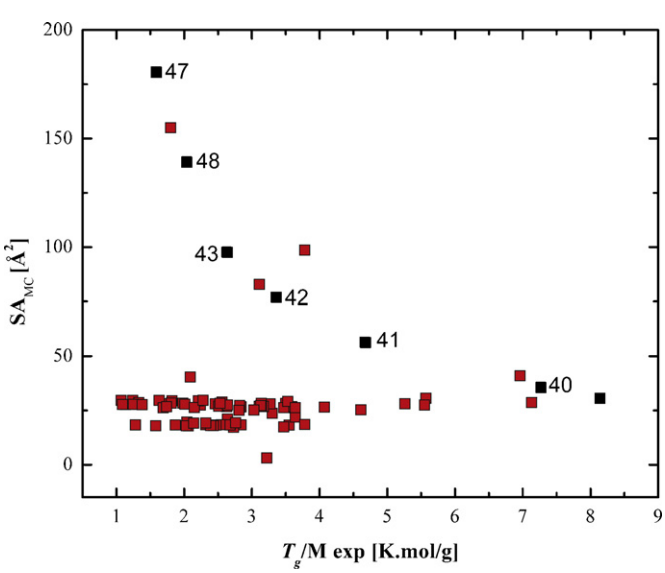

Fig. 6. Plot of SA_{MC} values versus experimental T_g/M values. Polyoxides are highlighted.

3.4. Number of rotatable bonds (RBN)

The number of rotatable bonds (RBN) is a well-known constitutional descriptor [42] that is calculated on the entire trimer (see Section 2.3.2). It is the number of bonds allows free rotation around themselves. They are defined as any single bond, not in a ring, bound to a nonterminal heavy atom (i.e., nonhydrogen atoms). Amide C—N bonds are excluded from the count because of their high rotational energy barrier. In the case of polymers, RBN seems to be a good indicator of the side group length. For example, RBN is unable to differentiate polymers with short side chains, like #2 (Table 1), from those either with an aromatic group like #14, #15, #16, #17, #18, #19, #20 and #21, or with an aliphatic ring like #4 and #5, but it can distinguish polymers that have extended length side chains like those in #49, #63, #67, #70, #76, #77, #26, #27, #50, #55, #73, #45, #53, #51, #52, #46, and #54. Hence, this descriptor is useful for incorporating information related to the length of the side chain (number of methylene groups).

According to RBN values, in the case of the repeating unit has an equal amount of matter, an inverse relationship between RBN and T_g/M is generally observed (Fig. 7). This effect is clearly evidenced in isomers of acrylates, such as #26 and #9 whose side groups are structurally different: n-butyl presents higher RBN than sec-butyl, respectively (Table 7). These molecules are not differentiated by the remaining descriptors of the model. A similar situation occurs in acrylates and methacrylates: even though they are not isomers, the longer methylene chain length of side group, the higher RBN and the lower T_g/M are, as shown by the experimental evidence described by Van Krevelen [51]. In general, the RBN descriptor fails to describe T_g/M's behavior in the case of molecules with few bonds that rotate freely round themselves and also with many substituents because the calculation does not consider these bonds

Table 6
SA_{MC}, RBN, M_{SC} and T_g/M values corresponding to some isomers of the dataset.

M	N	Name of polymer	SA_{MC}	M_{SC}	RBN	T_g/M exp
96	#4	Poly(cyclopentylethylene)	26.163	69.126	7	3.63
98	#56	Poly(1-heptene)	27.503	71.142	16	2.24
100	#8	Poly(ethyl acrylate)	27.872	73.072	13	2.51
100	#32	Poly(methyl methylacrylate)	18.551	74.079	10	3.78
110	#5	Poly(cyclohexylethylene)	23.718	83.153	7	3.30
110	#82	Poly(3-cyclopentyl-1-propene)	25.216	83.153	10	3.03
114	#58	Poly(n-propyl acrylate)	27.878	87.098	16	2.01
114	#33	Poly(ethyl methylacrylate)	18.363	88.106	13	2.84
128	#26	Poly(n-butyl acrylate)	28.081	101.125	19	1.71
128	#81	Poly(n-propyl methacrylate)	18.081	102.133	16	2.39
142	#70	Poly(3-pentyl acrylate)	28.386	115.152	19	1.81
142	#68	Poly(sec-butyl methacrylate)	18.46	116.16	16	2.32
142	#37	Poly(tert-butyl methylacrylate)	18.339	116.16	13	2.68
170	#53	Poly(n-heptyl acrylate)	27.788	143.206	28	1.25
170	#73	Poly(n-hexyl methacrylate)	17.956	144.214	25	1.58
198	#54	Poly(n-nonyl acrylate)	27.756	171.26	34	1.09
198	#52	Poly(n-octyl methylacrylate)	18.328	172.268	31	1.28

M is the repeating unit mass, N is the number of molecule.

Table 7
SA_{MC}, RBN, M_{SC} and T_g/M values corresponding to some isomers of the acrylates.

M	N	Name of polymer	SA_{MC}	M_{SC}	RBN	T_g/M exp
128	#25	Poly(tert-butyl acrylate)	28.039	101.125	13	2.46
128	#9	Poly(sec-butyl acrylate)	28.334	101.125	16	1.98
128	#26	Poly(n-butyl acrylate)	28.081	101.125	19	1.71

M is the repeating unit mass, N is the number of molecule.

when they are attached to a terminal heavy atom, like in #29, #62, #69, #61, #65, #78, #39 and #30.

3.5. Side chain mass (M_{SC})

In general, it can be observed that the greater the mass of the side chain, the lower the T_g/M values (Fig. 8), which is a trend that is similar to the one of the remaining model descriptors. Indeed, this descriptor provides relevant information because the mass of the side chain suffers more variation than the main chain in the dataset. Therefore, it will be very important in T_g/M ratio. It was also observed that, although there is some correlation between RBN and M_{SC}, each one provides particular and Supplementary information and both are relevant for the model, as it is demonstrated in the next section.

3.6. Evaluation of variable relevance

Apart from statistical measures (Section 3.1), the importance of the input variables to the generated ANN was assessed by removing the ith input variable, training the network without it and evaluating the resulting model by: R^2, MAE, MSE, RMS (root mean square) and MRE. The metrics were compared with the reference values obtained globally for the complete model (Table 8). When indexes denounced a significant deterioration, it could be concluded that the presence of the associated ith input variable was mandatory for the model. When indexes enhance or remains similar to

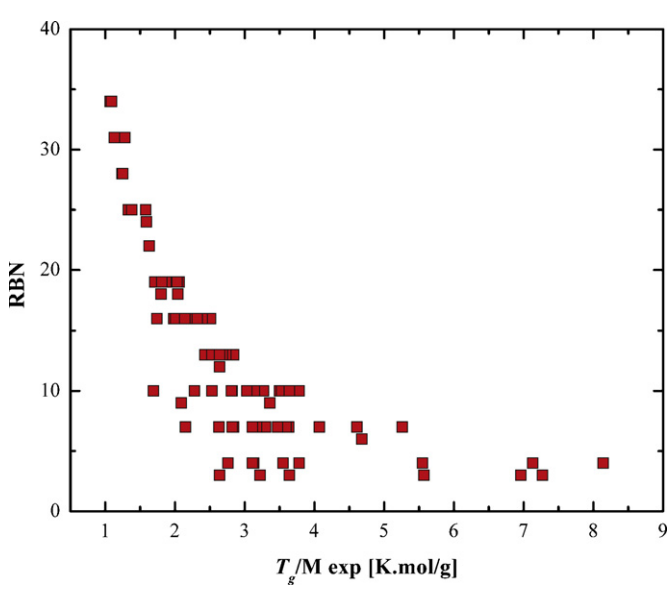

Fig. 7. Plot of RBN values versus experimental T_g/M values.

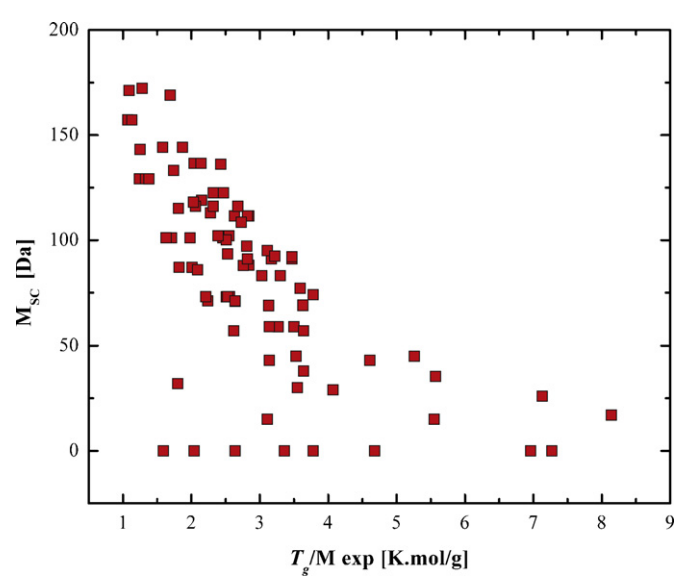

Fig. 8. Plot of M_{SC} values versus experimental T_g/M values.

Table 8
Statistical metrics for the input variables assessment.

Model	Set	%	R^2	MAE	MSE	RMS	MRE
DS1	Training	50	0.953	0.2	0.06	0.25	7.99
	Test	23	0.964	0.26	0.17	0.41	8.23
	Validation	27	0.933	0.28	0.22	0.47	9.81
DS1−{RBN}	Training	50	0.837	0.39	0.22	0.47	14.2
	Test	23	0.846	0.6	0.71	0.84	18.11
	Validation	27	0.84	0.4	0.25	0.5	14.42
DS1−{M_{SC}}	Training	50	0.751	0.41	0.34	0.58	14.54
	Test	23	0.796	0.63	1	1	17.3
	Validation	27	0.754	0.45	0.46	0.68	15.72
DS1−{SA_{MC}}	Training	50	0.744	0.37	0.35	0.59	12.11
	Test	23	0.863	0.53	1.01	1	13.04
	Validation	27	0.838	0.36	0.31	0.56	11.84

DS1−{RBN} is data splitting 1 minus RBN, DS1−{M_{SC}} is data splitting 1 minus M_{SC}, and DS1−{SA_{MC}} is data splitting 1 minus SA_{MC}.

original model, the *i*th input variable should be taken out. If the values are better, the variable is affecting the model negatively. In turn, when they are similar, the variable seems redundant.

From the results (Table 8), it is evident that all the input variables play an important role in the model since none of them neither reach nor overcome original-model performance.

4. Conclusions

In the present work a 3-descriptors QSPR model to predict T_g/M in high molecular weight polymers was proposed, obtaining a very good prediction performance. A novel set of 26 descriptors was proposed. In order to supplement the information of these new descriptors, also 635 classic descriptors were calculated for the entire trimer and a variable selection method was applied to whole pool of descriptors (new and classic ones) obtaining the better 3-descriptors model (SA_{MC}, M_{SC}, RBN), where the first two belong to the new ones. These three descriptors were use as input in an artificial neural network to generate the prediction model for T_g/M mentioned above.

The new proposed descriptors were defined on structural fragments corresponding to the main and side chains, based on the middle repeating unit of a trimeric structure. The easiness and versatility to calculate these descriptors constitute two of the main advantages of this strategy. Firstly, this operation can be automated since the definition of these descriptors is unambiguous. Secondly, this approach is independent of the type of atoms and atomic groups that constitute the polymer, therefore applicable to any type of polymer families. These new chain descriptors have a clear physicochemical interpretation since they capture genuine characteristics of the polymeric structure and they are comprehensible. It is remarkable the expression ability of new approach as 2 out of the 3 selected descriptors belongs to the new ones.

The resulting model presented a low number of descriptors (three) y statistical metrics comparable to the best ones reported in the literature so far, thus demonstrating the quality of the model for predicting T_g/M. A detailed structural explanation of the model descriptors and its relation to the property studied was presented. As a result, a model enriched with the underlying physicochemical knowledge of the studied phenomenon was obtained. Besides, the relevance of the input variables in the prediction model was also judged and all of them proved to be necessary.

These results are promising since this type of novel approach can be projected to the study of other properties of complex molecules, such as high molecular weight polymers, by providing a physicochemical support in addition to giving a statistical backing.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2012.04.006.

References

[1] A.R. Katritzky, S. Sild, V. Lobanov, M. Karelson, Quantitative structure–property relationship (QSPR) correlation of glass transition temperatures of high molecular weight polymers, Journal of Chemical Information and Computer Science 38 (1998) 300–304.
[2] 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.
[3] J.M. Barton, Relation of glass transition temperature to molecular structure of addition copolymers, Journal of Polymer Science Part C 30 (1970) 573–597.
[4] W.A. Lee, Calculation of the glass transition temperatures of polymers. Part I. Homopolymers and copolymers with alkyl side chains, Journal of Polymer Science Part A2 8 (1970) 555–570.
[5] H.G. Weyland, P.J. Hoftyzer, D.W. Van Krevelen, Prediction of the glass transition temperature of polymers, Polymer 11 (1970) 79–87.
[6] D.W. Van Krevelen, Properties of Polymers, second ed., Elsevier, Amsterdam, 1976.
[7] D.R. Wiff, M.S. Altieri, I.J. Goldfarb, Predicting glass transition temperatures of linear polymers: random copolymers, and cured reactive oligomers from chemical structure, Journal of Polymer Science Polymer Physics Edition 23 (1985) 1165–1176.
[8] J. Bicerano, Prediction of Polymer Properties, second ed., Marcel Dekker, New York, 1996.
[9] B.G. Sumpter, D.W. Noid, Neural networks and graph theory as computational tools for predicting polymer properties, Macromolecular Theory and Simulations 3 (1994) 363–378.
[10] S.J. Joyce, D.J. Osguthorpe, J.A. Padgett, G.J. Price, Neural network prediction of glass-transition temperatures from monomer structure, Journal of the Chemical Society, Faraday Transactions 91 (1995) 2491–2496.
[11] C.C. Cypcar, P. Camelio, V. Lazzeri, L.J. Mathias, B. Waegell, Prediction of the glass transition temperature of multicyclic and bulky substituted acrylate and methacrylate polymers using the energy, volume, mass (EVM) QSPR model, Macromolecules 29 (1996) 8954–8959.
[12] P. Camelio, V. Lazzeri, B. Waegell, C. Cypcar, L.J. Mathias, Glass transition temperature calculations for styrene derivatives using the energy, volume, and mass model, Macromolecules 31 (1998) 2305–2311.
[13] 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, Computational and Theoretical Polymer Science 8 (1998) 311–321.
[14] B.E. Mattioni, P.C. Jurs, Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks, Journal of Chemical Information and Computer Science 42 (2002) 232–240.
[15] C.Z. Cao, Y.B. Lin, Correlation between the glass transition temperatures and repeating unit structure for high molecular weight polymers, Journal of Chemical Information and Computer Science 43 (2003) 643–650.

[16] 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.
[17] C. Duce, A. Micheli, A. Starita, M.R. Tiné, R. Solaro, Prediction of polymer properties from their structure by recursive neural networks, Macromolecular Rapid Communications 27 (2006) 711–715.
[18] X. Yu, X. Wang, X. Li, J. Gao, H. Wang, Prediction of glass transition temperatures for polystyrenes by a four-descriptors QSPR model, Macromolecular Theory and Simulations 15 (2006) 94–99.
[19] 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, European Polymer Journal 43 (2007) 989–995.
[20] X. Yu, B. Yi, X. Wang, Z. Xie, Correlation between the glass transition temperatures and multipole moments for polymers, Chemical Physics 332 (2007) 115–118.
[21] X.L. Yu, B. Yi, X.Y. Wang, Prediction of the glass transition temperatures for polymers with artificial neural network, Journal of Theoretical and Computational Chemistry 7 (2008) 953–963.
[22] L.W. Ning, Artificial neural network prediction of glass transition temperature of fluorine-containing polybenzoxazoles, Journal of Materials Science 44 (2009) 3156–3164.
[23] W. Liu, C. Cao, Artificial neural network prediction of glass transition temperature of polymers, Colloid and Polymer Science 287 (2009) 811–818.
[24] M.G. Koehler, A.J. Hopfinger, Molecular modelling of polymers: 5. Inclusion of intermolecular energetics in estimating glass and crystal-melt transition temperatures, Polymer 30 (1989) 116–126.
[25] R. García-Domenech, J.V. de Julian-Ortiz, Prediction of indices of refraction and glass transition temperatures of linear polymers by using graph theoretical indices, Journal of Physical Chemistry B 106 (2002) 1501–1507.
[26] J. Gao, X. Wang, X. Li, X. Yu, H. Wang, Prediction of polyamide properties using quantum-chemical methods and BP artificial neural networks, Journal of Molecular Modeling 12 (2006) 513–520.
[27] W.Q. Liu, P.G. Yi, Z.L. Tang, QSPR models for various properties of polymethacrylates based on quantum chemical descriptors, The QSAR & Combinatorial Science 25 (2006) 936–943.
[28] J.F. Dai, S.L. Liu, Y. Chen, C.Z. Cao, A quantitative structure–property relationship study on the glass transition temperature of polyacrylates, Acta Polymerica Sinica 3 (2003) 343–347.
[29] S. Mallakpour, M. Hatami, H. Golmohammadi, Prediction of inherent viscosity for polymers containing natural amino acids from the theoretical derived molecular descriptors, Polymer 51 (2010) 3568–3574.
[30] HyperChem™, Molecular Modeling System, Release 8.0.7 for Windows, Hypercube, Inc., Gainesville, USA, 2009, http://www.hyper.com/ (accessed 21.02.12).
[31] W. Hasel, T.F. Hendrickson, W.C. Still, A rapid approximation to the solvent accessible surface areas of atoms, Tetrahedron Computer Methodology 1 (1988) 103–116.
[32] W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, Journal of the American Chemical Society 112 (1990) 6127–6129.
[33] N. Bodor, Z. Gabanyi, C.K. Wong, A new method for the estimation of partition coefficient, Journal of the American Chemical Society 111 (1989) 3783–3786.
[34] A. Gavezotti, The calculation of molecular volumes and the use of volume analysis in the investigation of structured media and of solid-state organic reactivity, Journal of the American Chemical Society 105 (1983) 5220–5225.
[35] A.K. Ghose, A. Pritchett, G.M. Crippen, Atomic physicochemical parameters for three dimensional structure directed quantitative structure–activity relationships III: modeling hydrophobic interactions, Journal of Computational Chemistry 9 (1988) 80–90.
[36] V.N. Viswanadhan, A.K. Ghose, G.N. Revankar, R.K. Robins, Atomic physicochemical parameters for three dimensional structure directed quantitative structure–activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics, Journal of Chemical Information and Computer Science 29 (1989) 163–172.
[37] A. Leo, C. Hansch, D. Elkins, Partition coefficients and their uses, Chemical Reviews 71 (1971) 525–616.
[38] A.K. Ghose, G.M. Crippen, Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure–activity relationships. 2. Modeling dispersive and hydrophobic interactions, Journal of Chemical Information and Computer Science 27 (1987) 21–35.
[39] K.J. Miller, Additivity methods in molecular polarizability, Journal of the American Chemical Society 112 (1990) 8533–8543.
[40] DRAGON for Windows (Software for Molecular Descriptor Calculations), Version 5.5, Talete srl, Milan, Italy, 2007, http://www.talete.mi.it/ (accessed 21.02.12).
[41] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, second ed., Wiley-VCH, Weinheim, 2009.
[42] I.V. Tetko, P. Bruneau, H.W. Mewes, D.C. Rohrer, G.I. Poda, Can we estimate the accuracy of ADME-Tox predictions? Drug Discovery Today 11 (2006) 700–707.
[43] A.J. Soto, R.L. Cecchini, G.E. Vazquez, I. Ponzoni, Multi-objective feature selection in QSAR using a machine learning approach, The QSAR & Combinatorial Science 28 (2009) 1509–1523.
[44] STATISTICA (Data Analysis Software System), Version 8.0, StatSoft, Inc., Tulsa, USA, 2007, http://www.statsoft.com/ (accessed 21.02.12).
[45] J.G. Topliss, R.J. Costello, Chance correlations in structure–activity studies using multiple regression analysis, Journal of Medicinal Chemistry 15 (1972) 1066–1068.
[46] M. Chanda, Introduction to Polymer Science and Chemistry: A Problem-Solving Approach, CRC Press, Boca Raton, 2006.
[47] 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, Chemical Reviews 110 (2010) 5714–5789.
[48] S.W. Yin, Z. Shuai, Y.L. Wang, A quantitative structure–property relationship study of the glass transition temperature of OLED materials, Journal of Chemical Information and Computer Science 43 (2003) 970–977.
[49] Chapter 6: guidance on the principle of mechanistic interpretation, guidance document on the validation of (quantitative) structure–activity relationships [(QSAR)] models, in: Series on Testing and Assessment, No. 69, OECD Environment Health and Safety Publications, 2007.
[50] 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.
[51] D.W. Van Krevelen, Properties of Polymers, fourth ed., Elsevier, Amsterdam, 2009.