Materials Chemistry and Physics 172 (2016) 158–164

Contents lists available at ScienceDirect

Materials Chemistry and Physics

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

Encoding alternatives for the prediction of polyacrylates glass transition temperature by quantitative structure–property relationships

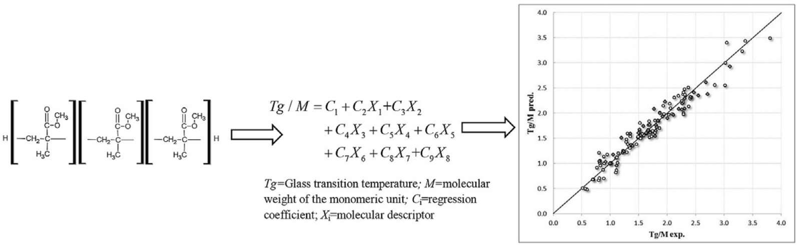
Andrew G. Mercader*, Pablo R. Duchowicz

Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA, UNLP, CCT La Plata-CONICET), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina

HIGHLIGHTS

- A QSPR model of the glass transition temperature of polyacrylates was build.
- The polymers structure encoding alternatives for this type of studies were explored.
- The validation procedures revealed very good predictive attributes by the model.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 20 July 2015
Received in revised form
3 December 2015
Accepted 23 January 2016
Available online 4 February 2016

Keywords:
Polymers
Glass transitions
Computer modeling and simulation
Computational techniques

ABSTRACT

The glass transition temperature, T_g, is one of the most important properties of amorphous polymers. The ability to predict the T_g value of a polymer prior to its synthesis it is of great value. For this reason we performed a predictive Quantitative Structure–Property Relationships (QSPR) analysis of T_g. The study explored the best way to encode the polymers structure for this type of studies, finding that the optimal option is using three monomeric units. The best linear model constructed from 126 molecular structures incorporated eight molecular descriptors and showed very good predictive ability, being a very simple and straight forward method for the prediction of T_g for polyacrylates since three dimensional descriptors were not used.

1. Introduction

The glass transition temperature, T_g, also known as the glass temperature or the glass–rubber transition temperature, is one of the most important properties of amorphous polymers [1].

As the temperature of a polymer drops below T_g, it behaves in an increasingly brittle manner. As the temperature rises above the T_g, the polymer becomes a rubber-like material. Thus, knowledge of T_g is essential in the selection of materials for various applications. In general, values of T_g well below room temperature define the domain of elastomers and values above room temperature define rigid structural polymers [1].

In the vicinity of T_g, a polymer experiences a sudden increase in the rate of molecular motions and, as a result, undergoes a series of conformational transformations. The torsional oscillations and/or rotations about most of the backbone bonds are activated, which causes a sharp increase in the free volume of the system as it is converted from the initial rigid (glassy) state to quasi-liquid state

* Corresponding author.
E-mail address: amercader@inifta.unlp.edu.ar (A.G. Mercader).

http://dx.doi.org/10.1016/j.matchemphys.2016.01.057
0254-0584/© 2016 Elsevier B.V. All rights reserved.

[2]. As a result of these processes, many physical properties of polymers change dramatically; for example, their coefficients of thermal expansion, heat capacities, and viscosities. The T_g is difficult to determine experimentally and predict theoretically because the transition takes place over a comparatively wide temperature range and is dependent on conditions such as the method of measurement, duration of the experiment, and pressure during the measurement [3,4]. The T_g is also very dependent on the structural (cross-linking, chain stiffness) [5], constitutional (additives, fillers, impurities) [6], and conformational (tacticity) features of polymers [1,4,7]. For these reasons, the discrepancies between reported values of T_g in the literature can be quite high [8].

Numerous researchers have attempted to predict T_g for polymers on the basis of Quantitative Structure Property Relationships (QSPR). According to the view of Katritzky et al., there are two kinds of approaches, the empirical and the theoretical [8]. Empirical methods correlate the studied property with other physical or chemical properties of the polymers, for instance, group additive properties (GAP) [1]. The GAP methodology is an entirely empirical approach, restricted to systems made merely of functional groups that have previously been investigated. It is an approximate method, since it fails to account the presence of neighboring groups or conformational influences. The most extensively referenced model made from theoretical estimations was proposed by Bicerano [4]; this regression model ($R = 0.9749$, $s = 24.65$ K) related the T_g with the solubility and the weighted sum of 13 structural parameters for a data set of 320 polymers; however the model was not tested on an external set, hence its validation was not assured.

Katritzky et al. [9] develop a mode with R^2 of 0.928 using 22 medium molecular weight polymers consisting of four parameters. Not presenting details about the way structures were encoded, only mentioning that tree to five monomeric units were used. Later on, Katritzky et al. [8] used CODESSA to predict the Tg for 88 linear homopolymers using five parameters and generated a QSPR model with a standard error of 32.9 K for Tg. In this case three monomeric units were used but no analysis was done to determine this number. In both these works, no external test sets were used; hence the models were not properly validated. Cao and Lin [10] tested the same set of 88 polymers using five parameters with clear physical meanings, calculated from individual repeating unit structures, finding a model with coefficient of determination of $R^2 = 0.9056$ and a standard error of 20.86 K. Once more, the model was not properly validated by an external test set.

Mattioni and Jurs [11] developed a 10-descriptor model using the structure of the monomer of 165 polymers, to predict T_g values using Artificial Neural Networks, the training set rms error was 10.1 K ($R^2 = 0.98$) and a prediction set (17 polymers) rms error of 21.7 K ($R^2 = 0.92$). In addition, an 11-descriptor model using one repeating unit from 251 different polymers, in this case, the training set rms error was 21.1 K ($R^2 = 0.96$) and a prediction set (25 polymers) rms error 21.9 K ($R^2 = 0.96$). Although the size of the prediction set is rather small, the results indicate that the use of the repeating unit instead of the monomer structure has a better predictive ability. In this article no further trials were done to attempt to determine the best number of repeating units to encode the structures.

A comprehensive neural network model with 28 descriptors was developed by Chen et al. [12] to predict T_g values of 6 randomly selected polymers from a database containing 71 polymers. The network was trained with the remaining 65 polymers, using descriptors calculated from individual repeating unit structures, and had training root mean square error of 17 K ($R^2 = 0.95$) and prediction average error of 17 K ($R^2 = 0.85$). Arriving at a presumably good model, however the number of test set polymers seems excessively low and the descriptors used excessively high, hence the predictivity of the model is uncertain.

A Support Vector Machine-based QSPR for the Prediction of Glass Transition Temperatures using 77 polymers was done by Yu [2]. Finding a model with root mean square (rms) errors for the training (38 polymers), validation (18 polymers) and prediction set (21 polymers) of 12.13, 15.58, and 16.22 K, respectively. Polymers were represented by one repeating unit end-capped by two hydrogen atoms, to calculate molecular descriptors.

An artificial neural network prediction of glass transition temperature using 113 polymers was done by Liu et al. [13], the final optimum neural network with produced a training set root mean square error (RMSE) of 11 K ($R = 0.973$) and a prediction set RMSE of 17 K ($R = 0.955$). To calculate the descriptors, the polymers were represented by their corresponding monomer.

As can be appreciated, none of the previous studies have evaluated the optimal number of monomeric units to represent the polymer structure in the prediction of T_g.

Recently a study using flexible descriptors successfully modeled a different property, the refractive index, using 234 structurally diverse polymers [14]. In this case the best found alternative was to encode the polymers with two repeating units.

In the case of polymer studies, it is not possible to calculate the molecular descriptors directly from the entire structure, since polymers possess very high molecular weights; moreover the size of the molecular chains may vary from different polymer preparations. Hence, the way to encode the molecules becomes a crucial part of a QSPR study involving polymers.

For that reason, the main objective of the present work is to study the best way to encode polymers in QSPR studies, in order to obtain reliable predictions based on a straight forward method. In order to do so, a dataset consisting of 126 polyacrylates was selected. Only polyacrylates were included in this study aiming to have a structurally similar set, and consequently producing more precise models.

2. Methods

2.1. Data sets

To carry out this study, a total of 126 polyacrylates with experimental T_g were taken from a published compilation [15], to our knowledge this set of molecules was not employed in this type of study before. Only the polyacrylates family was chosen aiming to produce a more specific and precise study. The experimental Tg values along with the SMILES structure representation can be found on Table S1. SMILES notation was chosen as a way of sharing the dataset with any interested reader, since it allows easily copying the text string and entering it in many chemical structure representation software. The data-set was divided into a training set of 84 and a test set of 42 polymers by applying a k-means cluster analysis [16], in order to have representative molecules of the structure diversity of the complete dataset in both training and test sets.

Following the procedure done by of Katritzky et al. [8] where T_g was divided by the molecular weight of the repeating unit (M), and after some preliminary tests that showed that using T_g/M presented better correlation results than using directly T_g; it was decided to use T_g/M for the study.

The experimental measure of T_g is a difficult task, which is revealed in the dispersion of experimental data for some polymers, complicating the correlation studies since they rely on the quality of the experimental dataset. When more than one value was found for the same polymer an average was used.

2.2. Molecular descriptors

As mentioned, it is not feasible to calculate descriptors directly for the entire polymer structures. Therefore, models consisting of repeating units, end-capped by hydrogen, were chosen as small, yet representative structures, to calculate the descriptors (Fig. 1 shows an example of the structure of poly(methyl methacrylate) encoded by three monomeric units).

In principle at least three units would be necessary to properly describe the way in which the monomers connect to each other. In addition, since several descriptors take into account the neighboring atoms and the way in which the structural information propagates through a molecule, having three connected monomeric units may serve as a representation of the way the structural information spreads thorough the polymer. To verify if this assumption is correct different trials using one, two and three monomeric units were done. Following the same reasoning adding four or five repeating units may further contribute to better represent the properties of the polymer, hence additional tests using four and five monomeric units were performed.

The increase in the complexity of the structures does not represent a problem in the descriptor calculation procedure, since even for the highest complex case of using five repeating units the calculation time is lower than 5 min (on a regular desktop PC) for the entire set of polymers. However, when using four or five monomeric units, and depending on the polymer, there might be limitations on the size of the structures on the free available version of the descriptor calculating software (Virtual Computational Chemistry Laboratory [17]) since it allows molecules with a maximum of 150 atoms.

A simple and straightforward descriptor calculation methodology was used. The structures of the compounds were written in SMILES notation and directly inputted in Dragon 5.0 [18] (available online at the Virtual Computational Chemistry Laboratory [17]) which calculates parameters of all types such as Constitutional, Topological, Geometrical, Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Functional Groups. Three dimensional descriptors along with quantum chemical and semi-empirical descriptors were excluded; since, as only a small representative part of the structure is used, its actual 3D disposition is unknown; this considerably simplifies the descriptor calculation procedure since SMILES notation can be used directly without the need of any previous optimization. Constant variables were excluded; the final descriptors pools contained 695, 701, 697, 689 and 682 descriptors for the cases of 1, 2, 3, 4 and 5 monomers unit, respectively.

2.3. Model search

The model search consists in finding an optimal subset **d** of *d* descriptor from a set **D**, containing *D* descriptors, with *d* << *D*, and with minimal standard deviation *S*,

$$S = \sqrt{\frac{1}{(N-d-1)}\sum_{i=1}^{N} res_i^2} \quad (1)$$

by means of the Multivariate Linear Regression (MLR) technique. In this equation *N* is the number of molecules in the training set, and res_i the residual for molecule *i*, is the difference between the experimental property ($\mathbf{p}$) and predicted property ($\mathbf{p_{pred}}$). More precisely, it is sought to obtain the global minimum of $S(\mathbf{d})$ where **d** is a point in a space of size $D!/[d!(D-d)!]$. A full search (FS) of optimal variables is impractical because it requires $D!/[d!(D-d)!]$ linear regressions. Therefore, an alternative method is necessary. The optimum set of descriptors was selected by using a new advanced version of the Enhanced Replacement Method (ERM) [19,20] as a search algorithm that produces linear regression QSAR models with results similar to the FS, but with much less computational work. This technique approaches the minimum of *S* by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of *d* descriptors $\mathbf{d} = \{X_1, X_2, \ldots, X_d\}$. The ERM [21] gives models with better statistical parameters than the Forward Stepwise Regression procedure [22], and the more elaborated Genetic Algorithms [23]. Details about the steps involved in the ERM algorithm are available elsewhere [24].

Amongst many other approaches to address this problem, principle component regression (PCR), partial least squares (PLS) and artificial neural networks (ANN) analyses provide highly predictive QSARs, however they are difficult to interpret for being abstract, and implement for not yielding an equation. A combination of GA and MLR has shown to produce simple, less sophisticated models with better performance on external testing set predictions than PLS [25]. In addition, on an extensive contrast work, ERM has shown to further improve the performance of the obtained models when compared to GA [23]; and since ERM provides the same type of models in terms of simplicity compared to GA, ERM was selected for this work.

In order to avoid common errors and pitfalls as presented in the review article by Le et al. [26], several test were carried out: the use of uninformative descriptors was checked through the correlation matrix (Table 1); possible overfitting was tested using a theoretical validation, and more importantly using a test set external validation; chance correlations were checked using a widely used *y*-randomization procedure [27]; and the domain of applicability of the models was informed using a Williams Plot (Fig. 3).

To theoretically validate the models, the well-known Leave-One-Out (*loo*) and the Leave-More-Out Cross-Validation procedures (*l-n%-o*) [28] were chosen, where *n%* accounts for the number of molecules removed from the training set. The number of cases for the removal of 20 random molecules was 1,000,000 in the case of Leave-More-Out. Calculations were done using the

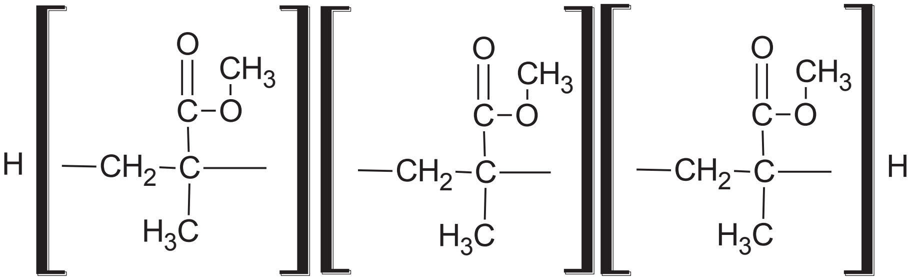


Fig. 1. Example of a trimeric repeating units for poly(methyl methacrylate).

Table 1
Correlation matrix for descriptors of Eq. (4) (N = 84).

	Se	MWC06	piPC01	IDDM	BEHm3	BELv8	nRCONHR	Neoplastic-50
Se	1	0.4941	0.7015	0.7981	0.0627	0.8763	0.0108	0.3524
MWC06		1	0.8184	0.8561	0.1667	0.3787	0.0056	0.2409
piPC01			1	**0.9527**	0.1607	0.7132	0.0374	0.3930
IDDM				1	0.1133	0.7330	0.0205	0.4121
BEHm3					1	0.0598	0.0064	0.1191
BELv8						1	0.0148	0.4352
nRCONHR							1	0.0563
Neoplastic-50								1

Bold face number indicate the highest correlation.

computational environment Matlab 5.0 (MathWorks, Natick, Massachusetts, U.S.A). The predictive ability of the model was further evaluated by $(r^2 - r^2{}_0)/r^2$, $(r^2 - r'^2{}_0)/r^2$, k and k′ [29,30].

The applicability domain (AD) for the QSAR models was explored in order to obtain reliable predictions for external samples. The AD is a theoretical region in the chemical space, defined by the model descriptors and modeled response, and thus by the nature of the chemicals in the training set, as represented in each model by specific molecular descriptors [31]. The AD can be characterized in various ways such as the leverage approach [32], which allows to verify whether a new chemical can be considered as interpolated and with reduced uncertainty or extrapolated outside the domain. If it is outside the model domain, a warning must be given. The leverage (h) is defined as [32]:

$$h_i = x_i\left(\mathbf{X}^T\mathbf{X}\right)^{-1} x_i^T \quad (i = 1, \ldots, M) \tag{2}$$

where x_i is the $1 \times d$ descriptor row-vector of compound i, M is the number of compounds in the dataset, and $\mathbf{X}$ is the $N \times d$ matrix of the training set (d is the number of model descriptors, and N is the number of training set samples). The leverage is suitable for evaluating the degree of extrapolation, its limit of normal values is set as $h^* = 3(N + 1)/M = 3(\Sigma h_i + 1)/M$, and a leverage greater than h^* for the training set means that the chemical is highly influential in determining the model, while for the test set, it means that the prediction is the result of substantial extrapolation of the model and may not be reliable.

The definition of the standardized residual (σ) for molecule i is:

$$\sigma_i = \frac{\text{res}_i}{S_{tr}} \tag{3}$$

where res_i is the residual of molecule i and S_{tr} is the standard deviation of the training set.

In order to visualize the AD of a QSAR model a Williams plot of standardized residuals (σ) *vs* leverage values (h) can be used to obtain an immediate and simple graphical detection of both the response outliers (Y outliers) and the structurally influential chemicals (X outliers) of a model.

3. Results and discussion

Using the ERM we searched the different pools of descriptors for models containing 1 to 10 molecular descriptors for the cases of the structures represented by 1–5 monomeric units; finding that the optimal number of descriptors for this dataset is 8. The optimal models obtained using T_g/M are presented in Table 2, where it can be seen that the best model is the model found using three monomeric units. This can be concluded by looking at the test set parameters, which are the most important since they reflect the predictability of the model and also expose if a true correlation between the experimental property and the molecular structure was found. It is clear that the models found by encoding the polymers using one monomeric unit are the worst, adding a second and a third monomeric unit improves the predictability of the models. These results corroborate that the way to represent the structure of the polymers require at least 3 monomers in order to properly indicate the way the monomeric unit connect with each other, which is also an additional proof that a true correlation between the structure and the measures property is present.

Table 2
Results of the best models found using different number of monomeric units to represent the polymers.

Monomers	d	S	R	FIT	S_{loo}	R_{loo}	S_{test}	R_{test}
1	8	0.1666	0.9742	9.436	0.1907	0.9661	0.2059	0.9303
2	8	0.1729	0.9727	8.913	0.1960	0.9649	0.1616	0.9554
3	**8**	**0.1697**	**0.9733**	**9.122**	**0.1999**	**0.9629**	**0.1515**	**0.9635**
4	8	0.1575	0.9770	10.638	0.1777	0.9707	0.1588	0.9548
5	8	0.1507	0.9790	11.673	0.1709	0.9729	0.1654	0.9525

Bold face numbers indicate the best model.

The model obtained using 4 monomers, although it might be comparable to the model with 3 monomers, has validation parameters of inferior quality. Adding a fifth monomer to the structure further weakens the validation parameters. If the possible previously mentioned limitations on the size of molecules by the descriptor calculating software, are also taken into account, it is clear that for the present data set adding more than three monomeric units is not advisable.

The model that better predicts the Tg/M using three monomers (third model of Table 2) is the following:

$$\begin{aligned} Tg/M = {} & 6.025(\pm 0.4) + 1.582 \times 10^{-2}\left(\pm 2 \times 10^{-3}\right)Se \\ & + 0.9433\ (\pm 0.1)\text{MWC06} + 2.1944(\pm 0.2)piPC01 \\ & - 3.811(\pm 0.2)IDDM - 0.2054\ (\pm 0.03)BEHm3 \\ & - 0.9992(\pm 0.2)BELv8 + 0.1802(\pm 0.02)nRCONHR \\ & + 0.4284(\pm 0.1)Neoplastic - 50 \end{aligned} \tag{4}$$

$$N = 84,\ R = 0.9733,\ S = 0.1697,\ FIT = 9.122,\ p < 10^{-4}$$
$$R_{loo} = 0.9629,\ S_{loo} = 0.1999,\ R_{l-20\%-o} = 0.7850,\ S_{l-20\%-o} = 0.4837$$
$$R_{TS} = 0.9635,\ S_{TS} = 0.1515$$

here, the standard errors of the regression coefficients are given in parentheses; p is the significance of the model, FIT the Kubinyi function, loo and l-20%-o stand for the Leave-One-Out and Leave-More-Out Cross Validation techniques respectively and *TS* stands for Test Set. Table 3 presents the meaning of the descriptors involved in Eq. (4). By looking at the regression coefficient of the test set, it can be seen that the predictive ability is comparable or better than most previously published models. The model obtained

Table 3
Symbols for molecular descriptors involved in the best model.

Molecular descriptor	Type	Description
Se	Constitutional indices	Sum of atomic Sanderson electronegativities (scaled on Carbon atom)
MWC06	Walk and path counts	Molecular walk count of order 6
piPC01	Path count	Molecular multiple path count of order 1
IDDM	Information indices	Mean information content on the distance degree magnitude
BEHm3	BCUT	Highest eigenvalue n. 3 of Burden matrix/weighted by atomic masses
BELv8	BCUT	Lowest eigenvalue n. 8 of Burden matrix/weighted by atomic van der Waals volumes
nRCONHR	Functional group counts	Number of secondary amides (aliphatic)
Neoplastic-50	Drug-like indices	Ghose-Viswanadhan-Wendoloski antineoplastic-like index at 50%

Italics indicate descriptors names.

by Mattioni and Jurs [11] apparently presents better results but using a smaller test set on a different data set.

To demonstrate that Eq. (4) are not the result of happenstance, we resorted to a widely used approach to establish the model robustness: the so-called *y*-randomization [27]. It consists of scrambling the experimental **p** property, so that activities do not correspond to the respective compounds. After analyzing 1,000,000 cases of y-randomization, the smallest *S* value obtained in this way was 0.5472, which is much larger than the one coming from the true calibration (0.1697). These results suggest that the model is robust, that its calibration is not a fortuitous correlation, and that a reliable structure–activity relationship was derived.

The plot of predicted by Eq. (4) *vs.* experimental Tg/M shown in Fig. 2 suggests that the 84 compounds from the training set and 42 from the test set tend to follow a straight line. The predicted activity given by Eq. (4) for the training and test sets are shown in Table S1. The Williams plot of the standardized residual *vs.* the leverages illustrated in Fig. 3 shows that most compounds lie within the AD of Eq. (4) and hence were calculated correctly, this is in line with the fact that an homologous series of compounds (polyacrylates) was used. Compounds **17** and **109** are X outliers of the training set reinforcing the model [32]; there are no compounds with a standardized residual higher than the limit (3σ) that can be considered outliers, compound **86** of the training set is the compound with the highest residual (2.8σ).

The correlation matrix of the model was presented in Table 1, descriptors *IDDM* and *piPC01* show a relevant degree of inter-correlation, however the calibration and validation results indicate that they are important for the prediction of the activity.

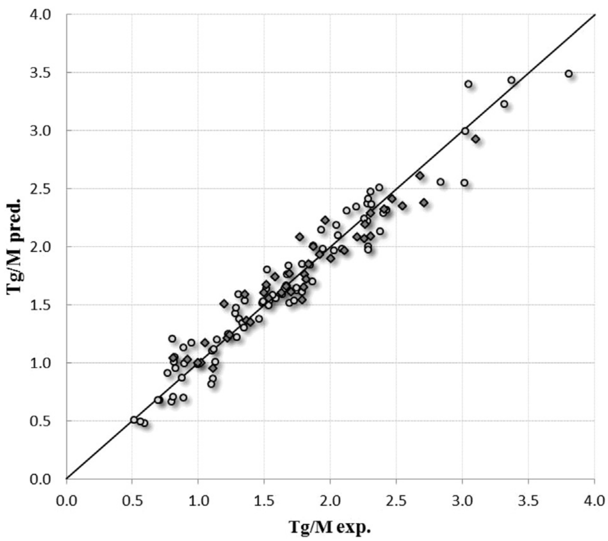


Fig. 2. Predicted (Eq. (4)) vs experimental Tg/M for the training (circles) and test (rhombus) sets.

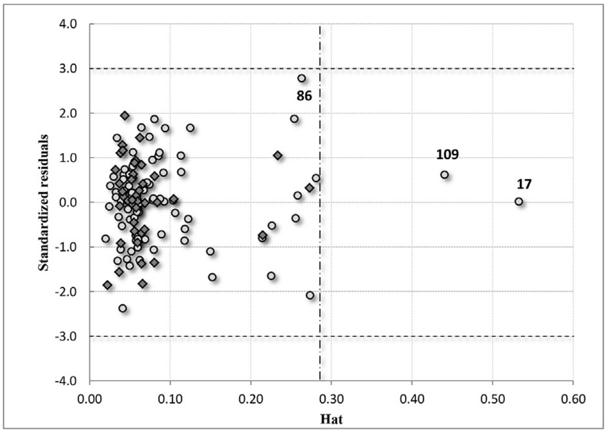


Fig. 3. Williams plot of the Eq. (4) showing the Application Domain for the training (circles) and test (rhombus) sets. The vertical dashed line indicates the limiting leverage h*.

The predictive power of the linear model is satisfactory as revealed by its stability upon the inclusion and/or exclusion of compounds, measured by the statistical parameter $R_{loo} = 0.9629$ (${R_{loo}}^2 = 0.9272$) and $R_{l-20\%-o} = 0.7850$ (${R_{l-20\%-o}}^2 = 0.6162$). As general rule $R_{l-n\%-o}$ (Q) should be higher than 0.71 ($Q^2 > 0.5$) to have a validated model [33,30].

The model was further validated by the following conditions [29,30]: ${R_{TS}}^2 = 0.9283 > 0.6$; k = 1.012; k′ = 0.9816 (0.85 < k or k′ < 1.15); $(r^2 - {r^2}_0)/r^2 = -0.0748 < 0.1$; $(r^2 - {r'^2}_0)/r^2 = -0.0728 < 0.1$.

The standardization of their regression coefficients of Eq. (4) allows assigning greater importance to the molecular descriptors that exhibit the largest absolute standardized coefficients [22]. In this case we have:

$$\begin{aligned} &IDDM(2.554) > piPC01(1.038) > MWC06(0.6832) \\ &> Se(0.6107) > BELv8(0.3439) > \mathrm{nRCONHR}(0.1829) \\ &> BEHm3(0.1518) > Neoplastic-50(0.1305) \end{aligned} \tag{5}$$

By looking at this order we can see that the most significant descriptor is the information index *IDDM*, followed by the Path Count descriptor *piPC01* and the Walk Count descriptor *MCWC06*.

Although a physical interpretation of the descriptors is normally not straight forward, the classes and some details of the most relevant descriptors appearing in Eq. (4) are given below.

Information indices are molecular descriptors calculated as information content of molecules, based on the calculation of equivalence classes from the molecular graph. Among them, the indices of neighborhood symmetry take into account also neighbor degree and edge multiplicity. They are calculated by applying formulas as the information content, which is a measure of the degree of diversity of the elements in a system. The information content of a system having n elements in a set.

$\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ is defined as:

$$I_C = \sum_{g=1}^{G} n_g log_2 n_g \tag{6}$$

where G is the number of different equivalence classes and n_g is the number of elements in the g^{th} class [34]. The descriptor *IDDM* is defined as the mean information content on the distance degree magnitude.

Path and Walk Counts are atomic and molecular descriptors obtained from a H-depleted molecular Graph, based on the counting of graph paths. The length of the path (the number of edges along the path) is called path order. To take into account multiple bonds and heteroatoms, weighted path counts can be calculated, either by introducing the weighting factors after the paths have been enumerated or by computing the weighted paths directly. Among other uses, molecular path codes can be employed to search for similarities among molecules [34]. They are related to the molecular brunching and size, and in general to the molecular complexity. The descriptor *piPC01* is defined as molecular multiple path count of order 1; and *MCWC06* as molecular walk count of order 6.

Constitutional indices are OD-descriptors, independent from molecular connectivity and conformations. In the case of *Se* is the sum of atomic electronegativities; and in the case of descriptor *nRCONHR* is calculated by counting the number of secondary aliphatic amides in the structure.

BCUT descriptors are obtained from the positive and negative eigenvalues of the adjacency matrix, weighting the diagonal elements with a type of atom weight. In the case of *BELv8* is weighted by the atomic volume; and in the case of *BEHm3* is weighted by the atomic mass.

4. Conclusions

In this paper we constructed a predictive QSPR model of the Tg/M based on 126 polyacrylates using eight molecular descriptors. The model can be used in a very straightforward manner since it is not based in 3D descriptors. The study showed that the optimal way to encode the polymers structures is to use three monomeric units. The best model exhibited great predictive ability established by theoretical and test set validations; and as could be appreciated it is of comparable or higher quality than most previously published models; it is advisable to use it specifically for polyacrylates since it was built with structurally similar polymers. We expect the proposed model to be a useful tool in the prediction of Tg activity, in a fast and costless manner, for any future studies that may require an estimation of this important property of polyacrylates.

Acknowledgments

The authors thank the National Research Council of Argentina (CONICET) and INIFTA (CONICET, UNLP) for financial support. PRD acknowledges the financial support from the National Research Council of Argentina (CONICET) PIP11220100100151 project and to Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.matchemphys.2016.01.057.

References

[1] D.W. Van Krevelen, Properties of Polymers, Elsevier, Amsterdam, 1990.

[2] X. Yu, Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers, Fibers Polym. 11 (5) (2010) 757–766, http://dx.doi.org/10.1007/s12221-010-0757-6.

[3] S. Krause, J.J. Gormley, N. Roman, J.A. Shetter, W.H. Watanabe, Glass temperatures of some acrylic polymers, J. Polym. Sci. Part A Polym. Chem. 3 (10) (1965) 3573–3586, http://dx.doi.org/10.1002/pol.1965.100031020.

[4] J. Bicerano, Prediction of Polymer Properties, New York, second ed., 1996.

[5] J.R. Potts, D.R. Dreyer, C.W. Bielawski, R.S. Ruoff, Graphene-based polymer nanocomposites, Polymer 52 (1) (2011) 5–25. http://dx.doi.org/10.1016/j.polymer.2010.11.042.

[6] K. Song, Y. Zhang, J. Meng, E. Green, N. Tajaddod, H. Li, M. Minus, Structural polymer-based carbon nanotube composite fibers: understanding the processing–structure–performance relationship, Materials 6 (6) (2013) 2543.

[7] P.-C. Ma, N.A. Siddiqui, G. Marom, J.-K. Kim, Dispersion and functionalization of carbon nanotubes for polymer-based nanocomposites: a review, Compos. Part A Appl. Sci. 41 (10) (2010) 1345–1367. http://dx.doi.org/10.1016/j.compositesa.2010.07.003.

[8] A.R. Katritzky, S. Sild, V. Lobanov, M. Karelson, Quantitative Structure–Property Relationship (QSPR) correlation of glass transition temperatures of high molecular weight polymers, J. Chem. Inf. Comput. Sci. 38 (2) (1998) 300–304, http://dx.doi.org/10.1021/ci9700687.

[9] A.R. Katritzky, P. Rachwal, K.W. Law, M. Karelson, V.S. Lobanov, Prediction of polymer glass transition temperatures using a general quantitative structure–property relationship treatment, J. Chem. Inf. Comput. Sci. 36 (4) (1996) 879–884, http://dx.doi.org/10.1021/ci950156w.

[10] C. Cao, Y. Lin, Correlation between the glass transition temperatures and repeating unit structure for high molecular weight polymers, J. Chem. Inf. Comput. Sci. 43 (2) (2003) 643–650, http://dx.doi.org/10.1021/ci0202990.

[11] B.E. Mattioni, P.C. Jurs, Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks, J. Chem. Inf. Comput. Sci. 42 (2) (2002) 232–240, http://dx.doi.org/10.1021/ci010062o.

[12] X. Chen, L. Sztandera, H.M. Cartwright, A neural network approach to prediction of glass transition temperature of polymers, Int. J. Intell. Syst. 23 (1) (2008) 22–32, http://dx.doi.org/10.1002/int.20256.

[13] W. Liu, C. Cao, Artificial neural network prediction of glass transition temperature of polymers, Colloid Polym. Sci. 287 (7) (2009) 811–818, http://dx.doi.org/10.1007/s00396-009-2035-y.

[14] P.R. Duchowicz, S.E. Fioressi, D.E. Bacelo, L.M. Saavedra, A.P. Toropova, A.A. Toropov, QSPR studies on refractive indices of structurally heterogeneous polymers, Chemom. Intell. Lab. Syst. 140 (0) (2015) 86–91. http://dx.doi.org/10.1016/j.chemolab.2014.11.008.

[15] A.A. Askadskii, Computational Materials Science of Polymers, Cambridge International Science Publishing, Cambridge, UK, 2003.

[16] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, Wiley-Interscience, New York, 2005.

[17] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, Virtual computational chemistry laboratory–design and description, J. Comput. Aided Mol. Des. 19 (6) (2005) 453–463, http://dx.doi.org/10.1007/s10822-005-8694-y.

[18] DRAGON. Release 5.0 Evaluation Version http://www.disat.unimib.it/chm.

[19] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Advances in the replacement and enhanced replacement method in QSAR and QSPR theories, J. Chem. Inf. Model. 51 (7) (2011) 1575–1581, http://dx.doi.org/10.1021/ci200079b.

[20] A. Lee, A.G. Mercader, P.R. Duchowicz, E.A. Castro, A.B. Pomilio, QSAR study of the DPPH radical scavenging activity of di(hetero)arylamines derivatives of benzo[b]thiophenes, halophenols and caffeic acid analogues, Chemom. Intell. Lab. Syst. 116 (0) (2012) 33–40. http://dx.doi.org/10.1016/j.chemolab.2012.03.016.

[21] A.G. Mercader, P.R. Duchowicz, F.M. Fernandez, E.A. Castro, Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories, Chemom. Intell. Lab. Syst. 92 (2008) 138–144.

[22] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley&Sons, New York, 1981.

[23] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Replacement method and enhanced replacement method *versus* the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories, J. Chem. Inf. Model. 50 (9) (2010) 1542–1548, http://dx.doi.org/10.1021/ci100103r.

[24] A.G. Mercader, P.R. Duchowicz, Enhanced replacement method integration with genetic algorithms populations in QSAR and QSPR theories, Chemom. Intell. Lab. Syst. 149 (Part A) (2015) 117–122. http://dx.doi.org/10.1016/j.chemolab.2015.10.007.

[25] A.K. Saxena, P. Prathipati, Comparison of MLR, PLS and GA-MLR in QSAR analysis, SAR QSAR Environ. Res. 14 (5–6) (2003) 433–445, http://dx.doi.org/10.1080/10629360310001624015.

[26] 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, http://dx.doi.org/10.1021/cr200066h.

[27] S. Wold, L. Eriksson, Statistical validation of QSAR results, in: Hvd Waterbeemd (Ed.), Chemometrics Methods in Molecular Design, vol. 2, VCH, Weinheim, 1995, pp. 309–318.

[28] D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Model. 43 (2003) 579–586.

[29] V. Ravichandran, S. Shalini, K. Sundram, A.D. Sokkalingam, QSAR study of substituted 1,3,4-oxadiazole naphthyridines as HIV-1 integrase inhibitors, Eur. J. Med. Chem. 45 (7) (2010) 2791–2797. http://dx.doi.org/10.1016/j.ejmech.2010.02.062.

[30] K. Roy, On some aspects of validation of predictive quantitative structure-activity relationship models, Expert Opin. Drug Discov. 2 (12) (2007) 1567–1577, http://dx.doi.org/10.1517/17460441.2.12.1567.

[31] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (5) (2007) 694–701, http://dx.doi.org/10.1002/qsar.200610151.

[32] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ. Health Perspect. 111 (10) (2003) 1361–1375.

[33] A. Golbraikh, A. Tropsha, Beware of q2!, J. Mol. Graph. Model. 20 (4) (2002) 269–276.

[34] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Vols. I & II, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009.