



Different encoding alternatives for the prediction of halogenated polymers glass transition temperature by quantitative structure–property relationships

Andrew G. Mercader, Daniel E. Bacelo & Pablo R. Duchowicz

To cite this article: Andrew G. Mercader, Daniel E. Bacelo & Pablo R. Duchowicz (2017): Different encoding alternatives for the prediction of halogenated polymers glass transition temperature by quantitative structure–property relationships, International Journal of Polymer Analysis and Characterization, DOI: [10.1080/1023666X.2017.1358847](https://doi.org/10.1080/1023666X.2017.1358847)

To link to this article: <http://dx.doi.org/10.1080/1023666X.2017.1358847>



Accepted author version posted online: 26
Jul 2017.
Published online: 26 Jul 2017.



Submit your article to this journal [↗](#)



Article views: 15



View related articles [↗](#)



View Crossmark data [↗](#)



Different encoding alternatives for the prediction of halogenated polymers glass transition temperature by quantitative structure–property relationships

Andrew G. Mercader^a, Daniel E. Bacelo^b, and Pablo R. Duchowicz^a

^aInstituto de Investigaciones Físicoquímicas Teóricas y Aplicadas, CCT La Plata-CONICET, UNLP, La Plata, Argentina;
^bDepartamento de Química, Facultad de Ciencias Exactas y Naturales, Universidad de Belgrano, Buenos Aires, Argentina

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 preceding its synthesis is of enormous value. For this reason it is of great value to perform a predictive quantitative structure–property relationships analysis of T_g , in this case a new set of halogenated polymers was used for this purpose. In addition, to corroborate our previous findings, the best way to encode the polymers structure for this type of studies was further tested finding that the optimal option is once more to use three monomeric units. The best linear model constructed from 153 molecular structures incorporated seven molecular descriptors and showed excellent predictive ability. Furthermore, the method showed to be very simple and straightforward for the prediction of T_g since three-dimensional descriptors are not required.

ARTICLE HISTORY




Submitted 11 June 2017
Accepted 20 July 2017

KEYWORDS

Computational techniques; computer modeling and simulation; glass transitions; halogenated polymers; QSPR

Introduction

The glass transition temperature, T_g , is one of the most important properties of amorphous polymers.^[1] When the temperature of a polymer gets lower than T_g , it starts behaving in an increasingly brittle manner. If the temperature is increased above the T_g , then the polymer recovers its rubber-like properties. Consequently, the knowledge of the T_g is crucial in the selection of the materials for a given application. In general, T_g values define the domain of elastomers or rigid structural polymers; having a T_g well below room temperature or above room temperature, respectively.^[1] In the vicinity of T_g , a polymer experiences an abrupt increase in the rate of molecular motion and, thus, undergoes a series of conformational transformations. The torsional oscillations and/or rotations of most backbone bonds are activated; this causes a sharp increase in the free volume of the system since it is converted from the initial rigid state to a quasi-liquid state.^[2] As a consequence, many physical properties of polymers change intensely; for example, their coefficients of thermal expansion, heat capacities, and viscosities. The T_g is difficult to determine experimentally and predict theoretically since the transition takes place over a wide temperature range and is dependent on many conditions, such as the measurement method, experiment duration, and pressure.^[3,4] T_g is also highly dependent on the structural (cross-linking, chain stiffness),^[5] constitutional (additives, fillers, impurities),^[6] and conformational (tacticity) characteristics of polymers.^[1,4,7] Consequently, the discrepancies between reported values of T_g can be quite high.^[8]

CONTACT Andrew G. Mercader  amercader@inifta.unlp.edu.ar; Pablo R. Duchowicz  pabloducho@gmail.com  Instituto de Investigaciones Físicoquímicas Teóricas y Aplicadas, CCT La Plata-CONICET, UNLP, Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina.

Numerous researchers have attempted to predict polymers T_g by quantitative structure–property relationships (QSPR). According to Katritzky et al.,^[8] there are two kinds of approaches, empirical and theoretical. Empirical methods attempt to find correlations between the studied property with other physical or chemical properties of polymers, for instance, group additive properties (GAP).^[11] The GAP methodology is a completely 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 effects. However, the most widely referenced theoretical model 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 of polymers, therefore its validation was not assured. Katritzky et al.^[9] developed a model with R^2 of 0.928 using 22 medium molecular weight polymers consisting of four parameters. Later on, Katritzky et al.^[8] used COmprehensive DEscriptors for Structural and Statistical Analysis (CODESSA) to predict the T_g for 88 linear homopolymers using five parameters and generated a QSPR model with a standard error of 32.9 K for T_g . 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 root mean square error (RMSE) was 10.1 K ($R^2 = 0.98$) and a prediction set (17 polymers) RMSE 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 RMSE was 21.1 K ($R^2 = 0.96$) and a prediction set (25 polymers) RMSE 21.9 K ($R^2 = 0.96$). A comprehensive neural network model with 28 descriptors was developed by Chen et al.^[12] to predict T_g values of six 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 RMSE 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 not certain. 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 RMSEs 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 and Cao,^[13] the final optimum neural network with produced a training set 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. 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 a recent work, our group has developed a QSPR T_g model based on 126 polyacrylates evaluating the optimal number of monomeric units to represent the polymer, since none of the previous studies have done so. An eight parameters optimal model found with training set $R = 0.9733$ and $S = 0.1697$, the test set values where $R = 0.9635$ and $S = 0.1515$. In addition, it was determined that the optimal encoding option is to use three monomeric units as representatives of the structure.^[15] 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. Accordingly, the main objective of the present work is to further corroborate the best way to encode polymers in QSPR studies in a completely different set of polymers. To do so, a data set consisting of 153 halogenated polymers was used, aiming to have a structurally similar nonetheless large set, to consequently produce more precise models.

Materials and methods

Data sets

In this study, a total of 153 halogenated polymers with experimental T_g were taken from a published compilation,^[16] to our knowledge this set of molecules was never used in this type of study before. Only the halogen containing polymers were chosen aiming to produce a more specific and precise study. The experimental T_g values and the SMILES structure representation can be found in Tables 1S and 2S. SMILES notation allows easily copying the text string to enter it in many chemical structure representation software. The data-set was divided into a training set of 102 and a test set of 51 polymers; it is known that doing this randomly does not lead to a rational selection, since both sets should have similar structure–property relationships having representative molecules of the structure diversity of the complete dataset. Consequently, the data set separation was performed by the balanced subsets method,^[17–19] based on the k -means cluster analysis.^[20] Following the procedure done by of Katritzky et al.^[8] and our previous work, T_g/M was used (where M stands for the molecular weight of the repeating unit). The experimental measurements of T_g is a problematic task, this is exposed in the dispersion of experimental data for some polymers, complicating the correlation studies as they rely on the quality of the experimental data set. When more than one value was informed for a given polymer an average was used.

Molecular descriptors

As mentioned in the introduction, it is not feasible calculating descriptors directly from 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 [Figure 1 shows an example of the structure of poly(2,5-dichlorostyrene) encoded by three monomeric units]. In theory, a minimum of three units would be necessary to properly describe the way in which the monomers connect to each other. In addition, because 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 might serve as a representation of the way the structural information spreads thorough the polymer. To verify this assumption, different trials using one, two and three monomeric units were completed. Following the same analysis adding four or five repeating units may further contribute to better represent the properties of the polymer, thus the use of four and five monomeric units were added to the tests. Increasing the number of monomeric units in the representative structure greatly increased the descriptor calculation time for polymer 153 (polyimide of 4,4'-diaminodiphenyloxide and 3,3',4,4'-tetracarboxydiphenylhexafluoroisopropylene), which presents a rather complex structure as is appreciated in Figure 2. The required calculation time on a desktop PC (CPU i7-4770 K 3.5 GHz with 8 GB DDR3 1600) is shown in Figure 3 where it can be seen the exponential growth of the calculation time with the number of monomeric units used. For the rest of the polymer the calculation time was not an issue since it was lower than 5 min even for five monomeric units. There is an additional restriction, when using four or five monomeric units, 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)^[21] as it allows molecules with a maximum of 150 atoms.

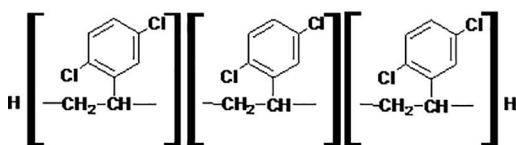


Figure 1. Example of a trimeric repeating units for poly(2,5-dichlorostyrene).

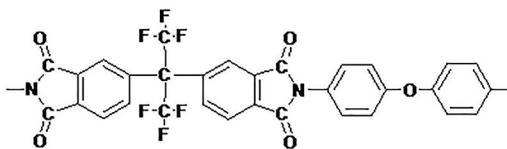


Figure 2. Monomeric unit structure of polymer 153.

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^[22] (available online at the Virtual Computational Chemistry Laboratory)^[21] 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 678, 680, 680, 678 and 692 descriptors for the cases of 1, 2, 3, 4 and 5 monomers unit, respectively.

Model search

A model search involves finding an optimal subset \mathbf{d} of d descriptor from a set \mathbf{D} , containing D descriptors, with $d \ll D$, and with minimal standard deviation S ,

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

by 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

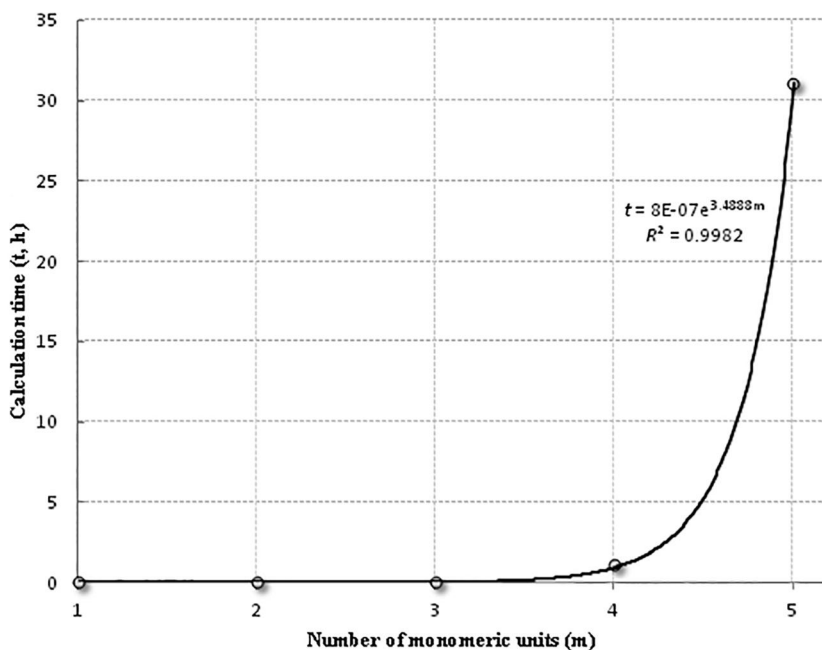


Figure 3. Number of monomeric units (m) in the structure vs calculation time (t) in hours. The exponential fit is shown in the graph.

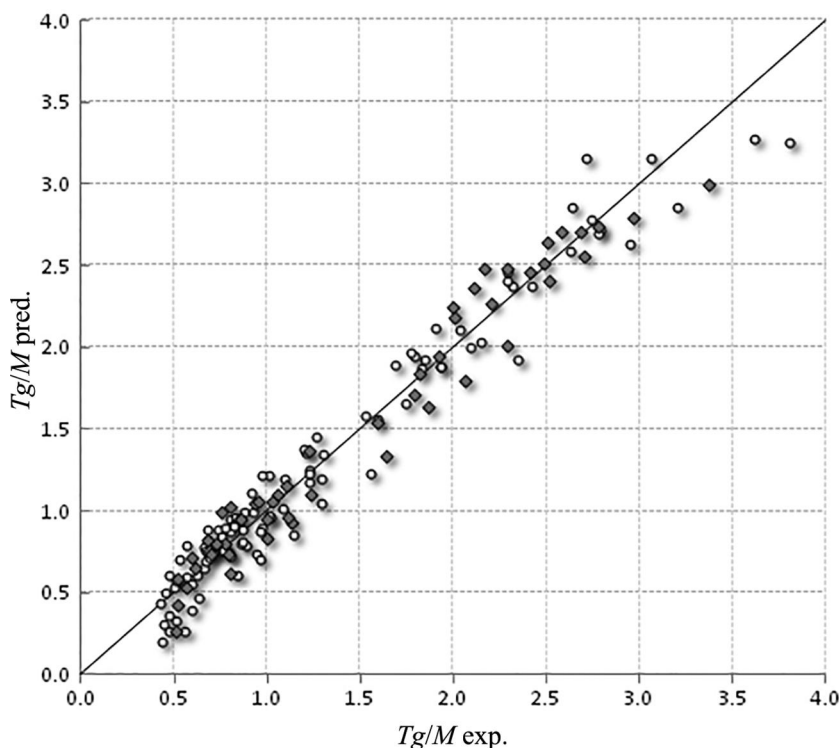


Figure 4. Predicted [Equation (4)] vs experimental T_g/M for the training (circles) and test (rhombus) sets.

property (\mathbf{p}) and predicted property (\mathbf{p}_{pred}). In other words, it is aimed to obtain the global minimum of $S(\mathbf{d})$ where \mathbf{d} is a point in a space of size $D!/[(d!(D-d)!)]$. A full search (FS) of optimal variables is impractical since it requires $D!/[(d!(D-d)!)]$ linear regressions. Therefore, an alternative method is necessary; for that reason, the optimum set of descriptors was selected using a new advanced version of the enhanced replacement method (ERM)^[23,24] as a search algorithm that produces linear regression QSPR models with results similar to the FS, nonetheless 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 $\mathbf{d} = \{X_1, X_2, \dots, X_d\}$.^[15] The ERM^[25] gives models with better statistical parameters than the forward stepwise regression procedure,^[26] and the more elaborated genetic algorithms.^[27] Details about the steps involved in the ERM algorithm are available elsewhere.^[28]

Amongst many other approaches to address this challenge, principle component regression, partial least squares (PLS) and artificial neural networks analyses provide highly predictive QSPRs, however they are difficult to interpret for being abstract, and to 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.^[29] In addition, on an extensive contrast work, ERM has shown to further improve the performance of the obtained models when compared to GA^[27]; and since ERM provides the same type of models in terms of simplicity compared to GA, ERM was selected for this work. To evade common errors and pitfalls as presented in the review article by Le et al.,^[30] numerous tests were performed: the use of uninformative descriptors was checked through the correlation matrix; 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 γ -randomization procedure^[31]; and the domain of applicability of the models was informed using a Williams plot (Figure 5).

Theoretically, validation of the models was done with the well-known leave-one-out (loo) and the leave-more-out cross-validation procedures (1- $n\%$ -o),^[32] 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 computational environment Matlab 5.0 (MathWorks, Natick, Massachusetts, USA). The predictive ability of the model was further evaluated by $(r^2 - r_0^2)/r^2$, $(r^2 - r_0'^2)/r^2$, k and k' .^[33,34] The applicability domain (AD) for the QSPR models was analysed 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.^[35] The AD can be characterized in various ways such as the leverage approach,^[36] which permits verifying 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)^[36] is defined as:

$$h_i = x_i(X^T X)^{-1} x_i^T \quad (i = 1, \dots, M) \quad (2)$$

where x_i is the $1 \times d$ descriptor row-vector of compound i , M is the number of compounds in the dataset, and 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 $h^* = 3(N + 1)/M = 3(\sum h_i + 1)/M$, 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:

$$\sigma_i = \frac{\text{res}_i}{S_{\text{tr}}} \quad (3)$$

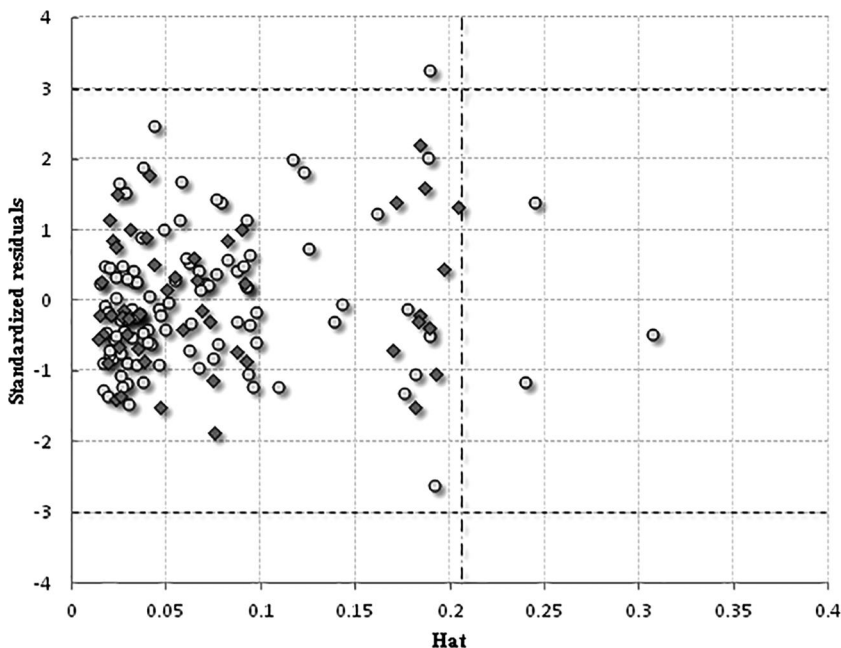


Figure 5. Williams plot of the Equation (4) showing the application domain for the training (circles) and test (rhombus) sets. The vertical dashed line indicates the limiting leverage h^* .

where res_i is the residual of molecule i and S_{tr} is the standard deviation of the training set. To visualize the AD of a QSPR model a Williams plot of standardized residuals (σ) vs leverage values (h) can be used to get an immediate and simple graphical detection of both: response outliers (Y outliers) and structurally influential chemicals (X outliers) of a model.

Results and discussion

Using the ERM we searched the five different pools of descriptors (structures represented by 1–5 monomeric units) for models containing 1–10 molecular descriptors; finding that the optimal number of descriptors for this dataset is 7. The optimal models obtained using T_g/M are presented in Table 1, where it can be seen that models using three and four monomeric units are very close in terms of the results. The statistical parameters of both models are comparable and some are better for model number 3 and some for model 4. It is clear that the model found by encoding the polymers using one monomeric unit is the worst, adding a second and a third monomeric unit improves the predictability of the models. These results verify that the way to represent the structure of the polymers requires at least three monomers 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.

The model obtained using four monomers, is comparable to the model with three monomers, however it presents a much higher correlation between the descriptors. Adding a fifth monomer to the structure deteriorates all statistical parameters. If the previously mentioned limitations on the size of molecules by the descriptor calculating software and the calculation time are taken into account, for the present data set adding using three monomeric units is advisable.

The model that better predicts the T_g/M using three monomers (third model of Table 1) is the following:

$$\begin{aligned} T_g/M = & 2.245(\pm 1) - 5.576(\pm 0.6)Me - 15.601(\pm 2)X4A \\ & - 3.587(\pm 0.4)X2Av + 1.998(\pm 0.3)BELp1 \\ & - 0.0279(\pm 0.007)SEigm + 24.76(\pm 0.9)VEe2 + 0.1434(\pm 0.04)nCR3X \end{aligned} \quad (4)$$

$$N = 102, R = 0.9786, S = 0.1703, FIT = 14.07, p < 2 \times 10^{-4} R_{loo} = 0.9723,$$

$$S_{loo} = 0.1934, R_{1-25\%-o} = 0.9504, S_{1-25\%-o} = 0.2711 R_{TS} = 0.9801, S_{TS} = 0.1595$$

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 1-25%-o stand for the leave-one-out and leave-more-out cross validation techniques respectively and TS stands for test set. Table 2 presents the meaning of the descriptors involved in Equation (4). By observing the regression coefficient of the test set, it can be seen that the predictive ability of the model is either comparable or better than most previously published models. To prove that Equation (4) is not the result of happenstance, we used y -randomization^[31] as a common method to establish the robustness of the model. It basically consists of scrambling the experimental p property, so that activities do not correspond to the respective compounds. After analysing 1,000,000 cases of y -randomization, the smallest S value

Table 1. Results of the best models found using different number of monomeric units to represent the polymers. Where C_{max} is the maximum correlation between any two descriptors in the model (please refer to Table 3). Boldface indicates the best results.

Monomers	d	S	R	FIT	S_{loo}	R_{loo}	S_{test}	R_{test}	C_{max}
1	7	0.1843	0.9753	12.124	0.2016	0.9703	0.2584	0.9513	0.9144
2	7	0.1739	0.9779	13.643	0.2038	0.9696	0.2436	0.9563	0.9384
3	7	0.1703	0.9786	14.071	0.1934	0.9723	0.1595	0.9801	0.6528
4	7	0.1722	0.9792	14.515	0.1971	0.9727	0.1485	0.9811	0.8744
5	7	0.1793	0.9750	11.999	0.1994	0.9690	0.1923	0.9735	0.9612

Table 2. Symbols for molecular descriptors involved in the best model.

Molecular descriptor	Type	Description
Me	Constitutional indices	Mean atomic Sanderson electronegativity (scaled on carbon atom)
X4A	Connectivity indices	Average connectivity index of order 4
X2Av	Connectivity indices	Average valence connectivity index of order 2
BELp1	BCUT	Lowest eigenvalue n. 1 of Burden matrix/weighted by atomic polarizabilities
SEigm	2D matrix	Spectral moment of order 1 from Barysz matrix weighted by mass
VEe2	2D matrix	Average coefficient of the last eigenvector from Barysz matrix weighted by Sanderson electronegativity
nCR3X	Functional group counts	number of CR3X

obtained in this way was 0.6693, which is much larger than the one coming from the true calibration (0.1703). These results suggest that the model is robust, that the calibration was not a fortuitous, and that a reliable structure–activity relationship was found.

The plot of predicted by Equation (4) vs experimental T_g/M presented in Figure 4 suggests that the 102 compounds from the training set and 51 from the test set tend to follow a straight line. The predicted values given by Equation (4) for the training and test sets are shown in Table 1S. The Williams plot of the standardized residual vs the leverages illustrated in Figure 5 indicates that most compounds rest within the AD of Equation (4) and hence were calculated correctly, this is in line with the fact that a restricted series of compounds (halogenated polymers) was used. Compounds **135**, **120** and **93** are training set X outliers reinforcing the model^[36]; chemical **89** has a standardized residual higher than 3σ and can be considered an outlier, however it belongs to the AD; this abnormal behavior could probably be attributed to wrong experimental data rather than to the molecular structure. The correlation matrix of the model was presented in Table 3, 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.

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 $R_{loo} = 0.9723$ ($R_{loo}^2 = 0.9454$) and $R_{1-25\%-o} = 0.9504$ ($R_{1-25\%-o}^2 = 0.9033$). As general rule $R_{1-n\%-o}(Q)$ should be higher than 0.71 ($Q^2 > 0.5$) to have a validated model.^[34,37] The model was further validated by the following conditions: $R_{TS}^2 = 0.9605 > 0.6$; $k = 0.9970$; $k' = 0.9945$ ($0.85 < k$ or $k' < 1.15$); $(r^2 - r_0^2)/r^2 = -0.0410 < 0.1$; $(r^2 - r_0^2)/r^2 = -0.0409 < 0.1$. The standardization of their regression coefficients of Equation (4) allows assigning greater importance to the molecular descriptors that exhibit the largest absolute standardized coefficients.^[26] In this case we have:

$$\begin{aligned} &VEe2(1.410) > Me(0.5505) > X2Av(0.3343) > X4A(0.3060) > BELp1(0.2308) > \\ &SEigm(0.1779) > nCR3X(0.08499) \end{aligned} \quad (5)$$

By looking at this order we can see that the most significant descriptor is the 2D matrix descriptor VEe2, followed by the constitutional index Me and the connectivity indices X2Av and X4A. Although a physical interpretation of the descriptors is normally not straight forward, the classes and some details of the most relevant descriptors appearing in Equation (4) are given below. The Barysz

Table 3. Correlation matrix for descriptors of Equation (4) ($N = 102$).

	Me	X4A	X2Av	BELp1	SEigm	VEe2	nCR3X
Me	1	0.3940	0.5343	0.3821	0.4940	0.1540	0.0059
X4A		1	0.4764	0.3850	0.4662	0.5185	0.0345
X2Av			1	0.2786	0.2849	0.2590	0.0185
BELp1				1	0.1224	0.6528	0.0388
Seigm					1	0.5792	0.2113
VEe2						1	0.2496
nCR3X							1

distance matrix D^Z is a weighted distance matrix accounting simultaneously for the presence of heteroatoms and multiple bonds in the molecule. It is defined as:

$$\langle D^Z \rangle_{ij} = 1 - \frac{Z_C}{Z_i} \quad (6)$$

were Z is the atomic number, Z_C obviously is the atomic number of carbon. For the case of VEE2 is the average coefficient of the last eigenvector from Barysz matrix weighted by Sanderson electronegativity. Constitutional indices are OD-descriptors, independent from molecular connectivity and conformations. In the case of Me is the mean of atomic electronegativities. Connectivity indices are topological molecular descriptors calculated from the vertex degree of the atoms in the H-depleted molecular graph. In the case of X2Av is the average valence connectivity index of order 2 and in the case of X4A average connectivity index of order 4.

Conclusion

In this paper we constructed a QSPR model of the T_g/M built by a new set of 153 halogenated polymers using seven molecular descriptors. The study showed that once more the optimal way to encode polymer structures is to use three monomeric units. The presented model exhibited great predictive ability established by theoretical and external set validations; showing to be of higher quality than most previously published models. The presented model can be used in a very straightforward manner since it does not require any structure optimization as is not based in 3D descriptors. Nevertheless, it is advisable to use it specifically for halogenated polymers to have a higher chance of using the model in its applicability domain, hence avoiding extrapolations. We assume that the proposed model will represent a useful tool for the prediction of T_g , in a rapid and costless manner, for any forthcoming studies that might need an estimation of this important property of halogenated polymers.

Funding

The authors thank the National Research Council of Argentina (CONICET) and INIFTA (CONICET, UNLP) for financial support. Pablo R. Duchowicz acknowledges the financial support from the National Research Council of Argentina (CONICET) PIP11220130100311 project and to Ministerio de Ciencia, Tecnología e Innovación Productiva for the electronic library facilities.

References

- [1] Van Krevelen, D. W. 1990. *Properties of Polymers*. Amsterdam: Elsevier.
- [2] Yu, X. 2010. Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers. *Fibers Polym.* 11:757–766.
- [3] Krause, S., J. J. Gormley, N. Roman, J. A. Shetter, and W. H. Watanabe. 1965. Glass temperatures of some acrylic polymers. *J. Polym. Sci. A Polym. Chem.* 3:3573–3586.
- [4] Bicerano, J. 1996. *Prediction of Polymer Properties*. New York: CRC Press.
- [5] Potts, J. R., D. R. Dreyer, C. W. Bielawski, and R. S. Ruoff. 2011. Graphene-based polymer nanocomposites. *Polymer* 52:5–25.
- [6] Song, K., Y. Zhang, J. Meng, E. Green, N. Tajaddod, H. Li, and M. Minus. 2013. Structural polymer-based carbon nanotube composite fibers: understanding the processing–structure–performance relationship. *Materials* 6:2543.
- [7] Ma, P.-C., N. A. Siddiqui, G. Marom, and J.-K. Kim. 2010. Dispersion and functionalization of carbon nanotubes for polymer-based nanocomposites: A review. *Compos. A Appl. Sci. Manuf.* 41:1345–1367.
- [8] Katritzky, A. R., S. Sild, V. Lobanov, and M. Karelson. 1998. Quantitative structure–property relationship (QSPR) correlation of glass transition temperatures of high molecular weight polymers. *J. Chem. Inf. Comput. Sci.* 38:300–304.

- [9] Katritzky, A. R., P. Rachwal, K. W. Law, M. Karelson, and V. S. Lobanov. 1996. Prediction of polymer glass transition temperatures using a general quantitative structure–property relationship treatment. *J. Chem. Inf. Comput. Sci.* 36:879–884.
- [10] Cao, C., and Y. Lin. 2003. Correlation between the glass transition temperatures and repeating unit structure for high molecular weight polymers. *J. Chem. Inf. Comput. Sci.* 43:643–650.
- [11] Mattioni, B. E., and P. C. Jurs. 2002. Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks. *J. Chem. Inf. Comput. Sci.* 42:232–240.
- [12] Chen, X., L. Sztera, and H. Cartwright. 2008. A neural network approach to prediction of glass transition temperature of polymers. *Int. J. Intell. Sys.* 23:22–32.
- [13] Liu, W., and C. Cao. 2009. Artificial neural network prediction of glass transition temperature of polymers. *Colloid. Polym. Sci.* 287:811–818.
- [14] Duchowicz, P. R., S. E. Fioressi, D. E. Bacelo, L. M. Saavedra, A. P. Toropova, and A. A. Toropov. 2015. QSPR studies on refractive indices of structurally heterogeneous polymers. *Chemom. Intell. Lab. Syst.* 140:86–91.
- [15] Mercader, A. G., and P. R. Duchowicz. 2016. Encoding alternatives for the prediction of polyacrylates glass transition temperature by quantitative structure–property relationships. *Mater. Chem. Phys.* 172:158–164.
- [16] Askadskii, A. A. 2003. *Computational Materials Science of Polymers*. Cambridge: Cambridge International Science Publishing.
- [17] Rojas, C., P. Tripaldi, and P. R. Duchowicz. 2016. A new QSPR study on relative sweetness. *Int. J. Quant. Struct.-Prop. Relat.* 1:78–93.
- [18] Rojas, C., P. R. Duchowicz, P. Tripaldi, and R. Pis Diez. 2015. Quantitative structure–property relationship analysis for the retention index of fragrance-like compounds on a polar stationary phase. *J. Chromatogr. A* 1422:277–288.
- [19] Rojas, C., P. R. Duchowicz, P. Tripaldi, and R. P. Diez. 2015. QSPR analysis for the retention index of flavors and fragrances on a OV-101 column. *Chemom. Intell. Lab. Syst.* 140:126–132.
- [20] Kaufman, L., and P. J. Rousseeuw. 2005. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley-Interscience.
- [21] Tetko, I., 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, and V. V. Prokopenko. 2005. Virtual computational chemistry laboratory – Design and description. *J. Comput. Aided Mol. Des.* 19:453–463.
- [22] DRAGON, Release 50 Evaluation Version. Available at: <http://www.disatunimibit/chm> (accessed March 7, 2016)
- [23] Mercader, A. G., P. R. Duchowicz, F. M. Fernández, and E. A. Castro. 2011. Advances in the replacement and enhanced replacement method in QSAR and QSPR theories. *J. Chem. Inf. Model.* 51:1575–1581.
- [24] Lee, A., A. G. Mercader, P. R. Duchowicz, E. A. Castro, and A. B. Pomilio. 2012. 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.* 116:33–40.
- [25] Mercader, A. G., P. R. Duchowicz, F. M. Fernandez, and E. A. Castro. 2008. Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. *Chemom. Intell. Lab. Syst.* 92:138–144.
- [26] Draper, N. R., and H. Smith. 1981. *Applied Regression Analysis*. New York: John Wiley & Sons.
- [27] Mercader, A. G., P. R. Duchowicz, F. M. Fernandez, and E. A. Castro. 2010. 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.* 50:1542–1548.
- [28] Mercader, A. G., and P. R. Duchowicz. 2015. Enhanced replacement method integration with genetic algorithms populations in QSAR and QSPR theories. *Chemom. Intell. Lab. Syst.* 149:117–122.
- [29] Saxena, A. K., and P. Prathipati. 2003. Comparison of MLR, PLS and GA-MLR in QSAR analysis. *SAR QSAR Environ. Res.* 14:433–445.
- [30] Le, T., V. C. Epa, F. R. Burden, and D. A. Winkler. 2012. Quantitative structure–property relationship modeling of diverse materials properties. *Chem. Rev.* 112:2889–2919.
- [31] Wold, S., and L. Eriksson. 1995. Statistical validation of QSAR results. In *Chemometrics Methods in Molecular Design*, ed. H. V. D. Waterbeemd. Weinheim: VCH, pp. 309–318.
- [32] Hawkins, D. M., S. Basak, and D. Mills. 2003. Assessing model fit by cross-validation. *J. Chem. Inf. Model.* 43:579–586.
- [33] Ravichran, V., S. Shalini, K. Sundram, and A. D. Sokkalingam. 2010. QSAR study of substituted 1,3,4-oxadiazole naphthyridines as HIV-1 integrase inhibitors. *Eur. J. Med. Chem.* 45:2791–2797.
- [34] Roy, K. 2007. On some aspects of validation of predictive quantitative structure–activity relationship models. *Expert Opin. Drug Discov.* 2:1567–1577.
- [35] Gramatica, P. 2007. Principles of QSAR models validation: Internal and external. *QSAR Comb. Sci.* 26:694–701.
- [36] Eriksson, L., J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, and P. Gramatica. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* 111:1361–1375.
- [37] Golbraikh, A., and A. Tropsha. 2002. Beware of q²! *J. Mol. Graphics Modell.* 20:269–276.