Chemometrics and Intelligent Laboratory Systems 140 (2015) 86–91

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

QSPR studies on refractive indices of structurally heterogeneous polymers

Pablo R. Duchowicz [a], Silvina E. Fioressi [b,*], Daniel E. Bacelo [b], Laura M. Saavedra [c], Alla P. Toropova [d], Andrey A. Toropov [d]

[a] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (CCT La Plata-CONICET, UNLP), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina
[b] Departamento de Química, Facultad de Ciencias Exactas y Naturales, Universidad de Belgrano, Villanueva 1324, CP 1426 Buenos Aires, Argentina
[c] Cátedra de Química Teórica y Computacional, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Calle 115 y 47, 1900 La Plata, Argentina
[d] IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, 20156, Via La Masa 19, Milano, Italy

ARTICLE INFO

Article history:
Received 15 August 2014
Received in revised form 14 November 2014
Accepted 15 November 2014
Available online 23 November 2014

Keywords:
QSPR theory
Polymer
Refractive index
Graph theory
Monte Carlo method
CORAL software

ABSTRACT

We developed a predictive Quantitative Structure–Property Relationship (QSPR) for the refractive indices of 234 structurally diverse polymers. The model involves a single molecular descriptor and a conformation-independent approach. The most appropriate polymer structure representation was investigated by considering 1–5 monomeric repeating units. The established equations were validated and tested through various well-known techniques, such as the use of an external test set of compounds, the Cross-Validation method, Y-Randomization and Applicability Domain, and finally a comparison was also performed to published results from the li terature. The developed QSPR could be useful for assisting the development of new polymeric materials.

1. Introduction

The most important optical property of a polymer is its refractive index (n). This value is directly related to optical, electrical and magnetic properties [1]. The manufacture of waveguides, optical films, optical fibers, and semiconductors among others is based in an extensive part on n. Polymers having low n in the range from 1.3 to 1.4 are useful, for example, as anti-reflective coatings for optical lenses. Transparent high refractive index polymers with $n \geq 1.5$ are used in optical applications demanding strong focusing power, because the material's lens effect increases with n. The refractive indices of organic polymers in the range from 1.3 to 1.7 can be accurately measured to four decimals at 23 °C through ASTM D 542-00 and ISO 489:1999 standard methods, using a V-prism refractometer. This measurement requires contact liquids to produce a planar contact between the sample and instrument's prism. Alterations in n due to chemical interaction with the contacting liquid must be avoided, limiting the number of contact liquids that can be used. Moreover, the liquid's n must be higher but not below one unit in the second decimal place when compared to the index of the polymer being measured. These standard methods are only suitable for isotropic materials. However, n of non-isotropic polymers may be determined using V-prism refractometer with slight modifications of the method, but with lower accuracy [2].

At present, the synthesis of new and more complex polymers with specific properties for diverse uses has greatly increased. Along with it, the range of chemical compounds involved in the polymerization process has become immense. In order to restrict the search for new materials with particular properties, prior to costly and time consuming synthesis, the use of predictive models is of key importance. The possibility of having a simple theoretical methodology to predict in a fast and accurate way the refractive indices of polymers are critical for the design of new and improved material generations [3–5].

During the last years, considerable efforts have been made to find out useful methodologies for predicting n [6], one of them being the QSPR theory [7–9]. Within the QSPR framework, the property of a chemical compound is completely determined by its molecular structure [10–16]. The interesting QSPR work of Bicerano [17] correlates the refractive index of 183 polymers with an eleven-descriptor model composed of topological and constitutional descriptors including connectivity indices and the total number of rotational degrees of freedom. The statistical quality involves an explained variance $R^2 = 0.95$ and a standard deviation $S = 0.0165$. In another study proposed by García-Domenech and de Julián-Ortiz [18], using the values of 121 amorphous linear polymers, a ten-descriptor model was obtained with the Best Multi Linear Regression method, achieving $R^2 = 0.96$ and $S = 0.015$.

* Corresponding author. Tel.: +54 11 47885400.
E-mail address: sfioressi@yahoo.com (S.E. Fioressi).

http://dx.doi.org/10.1016/j.chemolab.2014.11.008
0169-7439/

The set of graph-theoretical descriptors was calculated from the monomers and include Randic–Kier–Hall subgraph connectivity indices, their corresponding valence indices, topological charge indices, topological geometric indices, kappa indices, atom-type electrotopological state indices and the Wiener index. Although such models have a great accuracy, too many descriptors are involved which make them susceptible to overfitting and chance correlation [19].

Katritzky et al. [20] applied the Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA) software for modeling n in a set of 95 linear polymers. These authors developed a five-descriptor model, leading to a good correlation with $R^2 = 0.94$ and $S = 0.018$. The involved molecular descriptors are of two types: four quantum-chemical: the HOMO–LUMO energy gap, the AM1 heat of formation, the maximum nuclear repulsion for a carbon–hydrogen bond and the partial negative surface area (calculated from Zefirov's partial charges); one constitutional: the relative number of fluorine atoms. Xu et al. [21] proposed a four-descriptor equation based on the sum of valence degrees, the degree of unsaturation, the relative number of halogen atoms and the electrostatic attraction or hydrogen bond between the main chains. Their results from 121 polymers showed $R^2 = 0.93$ and $S = 0.018$. Recently, Astray and coworkers [22] presented a four-descriptor model with descriptors of the quantum-chemical type, obtained with Density Functional Theory (DFT) calculations at the B3LYP/6-31G(d) level. The set of descriptors calculated from the monomers in that study includes the energy of the lowest unoccupied molecular orbital, molecular average polarizability, heat capacity at constant volume, and the most positive net atomic charge on hydrogens. For a set of 95 polymers, they obtained $R^2 = 0.92$ and root mean square deviation ($RMSD$) of 0.023. Finally, it is to be noted that only the models established by García-Domenech and Astray were evaluated with external validation sets of compounds; validation is a crucial aspect in any QSPR study.

Every model that includes quantum-chemical descriptors usually involves a relatively difficult calculation of the optimum molecular geometry, involving high computational costs and long times. In this context, the conformation-independent 0D, 1D and 2D-QSPR methods emerge as an alternative approach for developing models based on constitutional and topological molecular features of compounds [23,24]. The exclusion of 3D-structural aspects also avoids problems associated with ambiguities, resulting from an incorrect computational geometry optimization due to the existence of compounds in various conformational states. Such kind of problems may also lead to the loose of predictive capability of the QSPR when applied for the prediction of an external test set of compounds.

In this work, we propose a flexible descriptor-based QSPR model [25] for the prediction of refractive index values, in a molecular set composed of 234 polymers with experimental information extracted from specialized books. In the realms of the approach used, the calculated flexible descriptor is a molecular descriptor which depends both on the molecular structure and the property under analysis (n), but does not explicitly depend on the 3D-molecular geometry. In previous QSPR studies, we have shown the importance of the methodology of flexible descriptors, which is able to provide models having a comparable or sometimes better quality to the ones found by searching the best descriptors in a pool containing thousands of 0D-3D descriptors [26–28]. Thereby, we investigated the most appropriate molecular structure representation for the flexible descriptor calculation, which can be done in different ways such as by using a chemical graph [29–31], using the Simplified Molecular Input Line Entry System (SMILES) [32–34], or with an hybrid representation which includes both graph and SMILES [35,36].

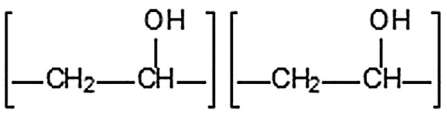

Fig. 1. Representative dimeric structure model for poly(vinyl alcohol).

2. Materials and methods

2.1. Experimental dataset

The high quality experimental refractive indices measured at 298 K on 234 polymer compounds were collected from two published compilations [17,37]. The n values range in the interval $[-9.91, 9.86]$, and the complete list of polymers studied here are included in Table 1S as Supplementary material. It is appreciated that the chemical domain analyzed is quite diverse, involving polyethylenes, polyacrylates, polymethacrylates, polystyrenes, polyether, polyoxides, polyamides and polycarbonates. The chemical groups of the side chains include halides, cyanides, carboxylates, acetates, amides, ethers, alcohols, hydrocarbon chains, aromatic rings, and non-aromatic rings.

2.2. Model development

2.2.1. Polymer structure model

Due to the high molecular weight of a polymer compound, it results impossible to directly calculate a molecular descriptor for the whole structure. Therefore, an alternative consists on proposing a representative structure model for the polymer, by means of resorting to a few number of repeating units (U). We used as basic though representative polymer structures the following cases: U, UU, UUU, UUUU and UUUUU. Fig. 1 offers an example for the case of poly(vinyl alcohol) in its dimeric representation. Whenever these model types are able to correlate relatively well the refractive indices, then they are assumed to be valid.

2.2.2. The flexible molecular descriptor definition

Several kinds of flexible molecular descriptors can be readily calculated with the CORAL freeware for Windows [38]. This software has been successfully applied previously in several QSPR studies and also for QSAR (Quantitative Structure–Activity Relationships) analyses [25].

As a first step, the repeating unit of each polymer has to be represented with SMILES notation, the chemical format used by CORAL. Table 2S includes the SMILES notations for the U-structure model. In the case of other structure models, the SMILES were easily prepared by coupling U components. For instance, poly(ethylene) having SMILES CC in the monomeric model results in CCCCCCCC in its tetrameric model. An advantage of working with CORAL in polymers is that hydrogen atoms can be avoided in the chemical structure for molecular descriptors calculation, which is not feasible in other software packages like Dragon [39] or CODESSA [40]. This is especially important for example for avoiding using terminal hydrogens in the monomeric polymer structure model, in order to differentiate in U-polymers like poly(ethylmethylene) and poly(1-methylethylene). Finally, the SMILES were provided as input to the CORAL program, together with the studied experimental property (n).

There are three different structural representation (SR) approaches available in the CORAL program, such as: i. using a chemical graph, like hydrogen-suppressed graph (HSG), hydrogen-filled graph (HFG) and graph of atomic orbitals (GAO); ii. using SMILES; and iii. using a hybrid representation which includes both graph and SMILES [25]. The SR used, ie. graph-based or SMILES-based, defines the number and types of structural attributes (local descriptors) that are able to take part in the QSPR analysis, and thus this specification defines the CORAL method. Therefore, one has to decide which particular combination of structural attributes should be considered the most appropriate for the modeling process.

In the graph approach of the HSG, HFG or GAO type, the structural attributes that can be used are the Morgan's extended connectivity indices of kth order for vertex (atom) Z ($^{k}EC_Z$, $k = 0$–3). It is noted that the index of zero-th order $^{0}EC_Z$ represents the vertex degree for

atom Z (number of neighbor atoms to Z in HSG), while the higher order indices ${}^{k}EC_Z$ are obtained through a recursive formula based on ${}^{0}EC_Z$ (see in Table 3S).

In the SMILES approach, the one-, two-, and three-element SMILES attributes ${}^{1}s_k$, ${}^{2}s_k$, and ${}^{3}s_k$, respectively, can be calculated. If a SMILES is a sequence of elements as 'ABCDE', then such structural attributes can be represented with Eqs. (1)–(3):

$$\text{'ABCDE'} \rightarrow \text{'A', 'B', 'C', 'D', 'E'} \left({}^{1}s_k\right) \tag{1}$$

$$\text{'ABCDE'} \rightarrow \text{'AB', 'BC', 'CD', 'DE'} \left({}^{2}s_k\right) \tag{2}$$

$$\text{'ABCDE'} \rightarrow \text{'ABC', 'BCD', 'CDE'} \left({}^{3}s_k\right). \tag{3}$$

In addition, the *NOSP*, *HALO*, *BOND* and *ATOMPAIR* attributes represent indices calculated according to the presence or absence of chemical elements: nitrogen, oxygen, sulfur, and phosphorus (*NOSP*); fluorine, chlorine, and bromine (*HALO*). The *BOND* is a mathematical function of the presence or absence of double (=), triple (#), or stereo chemical bonds (@ or @@). The *ATOMPAIR* is a mathematical function of the presence of seven chemical elements: F, Cl, Br, N, O, S, and P.

It is noteworthy that the way of searching for the most relevant structural attributes in a specific structural representation approach, in order to lead to the best statistical quality of the final model, is done in a stepwise fashion, in other words, first search for the best single attribute, then search for a second attribute that combines the best with the previous one, and then continue adding in this way the next attributes.

Within the CORAL framework, a QSPR model is obtained through a one-variable linear correlation between n and a properly defined flexible descriptor (*DCW*). The *DCW* descriptor is a linear combination (summation) of special coefficients, the so-called correlation weights (*CW*). A *CW* value is calculated for each type of structural attribute of the training set. The way for obtaining the *CW* values for all the structural attributes is based on the Monte Carlo simulation method, by optimizing a target function that depends upon the correlation coefficient (R) between n and the *DCW* descriptor (Table 3S).

The *DCW* flexible descriptor depends upon two positive integer values: the threshold value (T) and the number of epochs or iterations (N_{epochs}) used during the numerical optimization procedure. The appropriate selection of the threshold parameter avoids model overfitting, by classifying SMILES attributes into two categories: active and rare. The influence of rare attributes can be blocked by fixing their *CW* to zero. In this work, the rare attributes were the ones that take place in less than T polymers, while T was analyzed in the range from 0 to 5.

2.2.3. Model validation

The validation of the QSPR consists on testing its ability to predict the property for molecular structures not considered during the model development. The theoretical validation of the linear regression models is based on the popular validation criteria based on Cross Validation using Leave-One-Out (loo) and Leave-More-Out (ln%o, with n% being the percentile of molecules removed from the training set). The statistical parameters $R_{ln\ \%\ o}$ and $S_{ln\ \%\ o}$ (correlation coefficient and standard deviation of Leave-More-Out) measure the stability of the QSPR upon inclusion/exclusion of molecules. The number of cases for random data removal analyzed in this study is 100,000. According to the specialized literature, the loo explained variance (R^2_{loo}) should be greater than 0.5 for a validated model, although this is a necessary but not sufficient condition for its predictive power [13].

A more reliable validation was applied, that consists on using an external test set of structures. The 234 polymer compounds were ranked according to their refractive index values and every alternate compound was assigned to the training set (train), validation set (val) and test set (test). Each set thus included 78 compounds.

We used Y-Randomization [41] as a way of checking that the model does not result from happenstance and to avoid the development of fortuitous (chance) correlations. This technique consists on a permutation testing, and involves the same descriptors used in the models. New parallel models were developed on the basis of fit to randomly Y-data (Y-scrambling), and the process was repeated for a high number of iterations. After analyzing 10,000 cases of randomized response, the smallest standard deviation value obtained using this procedure (S^{rand}) has to be a higher (poorer) value than the one found by considering the true calibration (S).

2.2.4. Applicability domain

The applicability domain for the QSPR model was also explored, as not even a predictive model is expected to reliably predict the modeled property for the whole universe of molecules. The applicability domain is a theoretically defined area that depends on the molecular descriptor values and the experimental property analyzed [42]. Only the molecules falling within this applicability domain are not considered model extrapolations. One possible way to characterize the applicability domain is based on the leverage approach [43], which allows to verify whether a new compound can be considered as interpolated (with reduced uncertainty, reliable prediction) or extrapolated outside the domain (unreliable prediction). Each compound i has a calculated leverage value (h_i) and there exists a warning leverage value (h^*); Table 3S includes the definitions for h_i and h^*. When $h_i > h^*$ for a test set compound, then a warning should be given: it means that the prediction is the result of substantial extrapolation of the model and could not be treated as reliable.

3. Results and discussion

We performed the QSPR analysis by searching the best linear regression models on the training set of 78 polymer compounds, for each representative polymer structure model U, UU, UUU, UUUU and UUUUU. In doing so, the molecular structure representation that results the most appropriate for the flexible descriptor calculation has to be investigated, and also it has to be decided which structural attributes are the most efficient for each SR in order to take part during the flexible descriptor design. This means to select the appropriate CORAL method. In general, one optimizes the *DCW* flexible descriptor by increasing R^2_{train}, until the model starts to lose predictive capability in the validation set. This is the same situation that appears when one has to decide for the most predictive model among several multivariable linear regressions, having descriptors being searched in a pool containing thousands of them [44].

Table 1 contains a summary for the statistical quality of the best QSPR models found by trying different possible CORAL methods. It reveals that the best choice of SR for the polymer structures is a hybrid-approach that includes both graph (HFG type) and SMILES representations, with exception to the U-polymer that only considers a graph approach (HFG type). It is also observed that the best statistics (in terms of the R^2 and S parameters of the training, validation and test sets) results for the dimeric polymer structure, and keeps the

Table 1
The best QSPR models for the refractive index of polymers in their representative structures. The selected model appears in bold.

Polymer model	SR	Structural attributes	R^2_{train}	S_{train}	R^2_{val}	S_{val}	R^2_{test}	S_{test}
U	HFG	${}^{1}EC$, ${}^{2}EC$	0.95	0.016	0.92	0.021	0.84	0.030
UU	**HFG + SMILES**	${}^{2}EC$, ${}^{3}s_k$	**0.96**	**0.014**	**0.95**	**0.016**	**0.85**	**0.028**
UUU	HFG + SMILES	${}^{2}EC$, ${}^{3}s_k$	0.91	0.021	0.91	0.021	0.87	0.026
UUUU	HFG + SMILES	${}^{2}EC$, ${}^{3}s_k$	0.91	0.021	0.91	0.021	0.85	0.027
UUUUU	HFG + SMILES	${}^{2}EC$, ${}^{2}s_k$	0.90	0.023	0.90	0.022	0.85	0.028

polymer's size as small as possible, with percentages of explained variances of 96%, 95% and 85% in train, val and test, respectively. Finally, it is interesting to note that the trivial monomeric model also achieves an acceptable statistics, however, as it uses hydrogen-filled graphs, it is not able to differentiate between i.e. poly(ethylmethylene) and poly(1-methylethylene), and so we did not consider the U-model.

The statistics for the stepwise evolution of the dimeric model is presented in Table 2, where the first selected structural attribute is ^2EC, then the following ones are ^3s_k and *HALO* in that order. As we followed the common practice of keeping the model's size as small as possible (Ockham's razor), in order to avoid any fortuitous correlation, we did not consider more attributes in the *DCW* calculation because there is no further improvement. More complete details for the QSPR model established are the following:

$$n = 0.0019\, DCW_4 + 1.4740 \quad (4)$$

$N_{train} = 78,\ R^2_{train} = 0.96,\ S_{train} = 0.0137,\ F = 2033,\ p<10^{-4},\ o(3S) = 0$
$R^2_{loo} = 0.96,\ S_{loo} = 0.0142,\ R^2_{l20\%o} = 0.95,\ S_{l20\%o} = 0.0158,\ S^{rand} = 0.0590$
$N_{val} = 78,\ R^2_{val} = 0.95,\ S_{val} = 0.0160$
$N_{test} = 78,\ R^2_{test} = 0.85,\ S_{test} = 0.0280.$

Here, F is the Fisher parameter and $o(3S)$ indicates the number of outlier compounds having a residual (difference between experimental and calculated n) greater than three-times the S_{train} value. The parameters used for the DCW_4 calculation were $T = 1$ and $N_{epochs} = 7$.

A plot such as Fig. 2 for the predicted indices of refraction as function of the experimental values for the training, validation and test sets (numerical data are provided in Table 4S) reveals a tendency for the points to have a straight line trend. The dispersion plot of residuals (i.e. residuals as a function of predicted n) in Fig. 3 demonstrates that residuals tend to obey a random pattern around the zero line, suggesting that the assumption of the MLR technique is fulfilled. Eq. (4) has no outliers in the training set.

The approval of the internal validation process of Eq. (4) is evidenced by the stability of this equation upon the inclusion/exclusion of compounds from the training set, measured via the exclusion of one molecule at a time in leave-one-out ($R^2_{loo} = 0.96,\ S_{loo} = 0.0142$) and also by excluding 20% of the observations in leave-more-out (15 molecules, $R^2_{l20\ \%\ o} = 0.95,\ S_{l20\ \%\ o} = 0.0158$). A further step to assess the robustness of present equation is the application of Y-Randomization, demonstrating that $S_{train} < S^{rand}$ and thus the calibration does not result from happenstance and results in a valid structure–refractive index relationship. Eq. (4) also satisfies the necessary external validation conditions reported in [13,45]:

$R^2_{test} > 0.5$
$1 - R^2_0/R^2_{test} < 0.1$ and $1 - R'^2_0/R^2_{test} < 0.1$
$0.85 \le k \le 1.15$ and $0.85 \le k' \le 1.15$
$R^2_m > 0.5.$

The R^2_0, R'^2_0, k, k' and R^2_m parameters appear defined in Table 3S and, according to Table 5S, the proposed QSPR fulfills all these conditions.

An analysis of the applicability domain of Eq. (4) (with leverage values provided in Table 6S) suggests that all the compounds included in the test set belong to the applicability domain of the model ($h_i < h^*$)

Table 2
The stepwise search for finding the best QSPR model in a dimeric polymer. The selected model appears in bold.

Structural attributes	R^2_{train}	S_{train}	R^2_{val}	S_{val}	R^2_{test}	S_{test}
^2EC	0.94	0.018	0.90	0.023	0.84	0.029
^2EC, ^3s_k	**0.96**	**0.014**	**0.95**	**0.016**	**0.85**	**0.028**
^2EC, ^3s_k, *HALO*	0.97	0.011	0.96	0.014	0.86	0.027

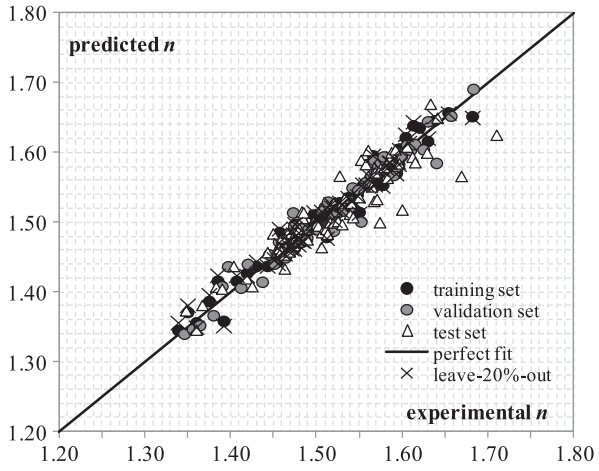

Fig. 2. Predicted (Eq. (4)) and experimental refractive indices for 234 polymer compounds.

with exception to **225**, poly(sulfone). After an exhaustive control of this compound in the source, we did not find any mistake in its experimental n value or its molecular structure. Hence, we assume that this particular behavior is due to the complexity of the data set, i.e. the structural heterogeneity of the molecules considered in this study. Thus, the predicted indices of refraction for all with exception to one test set polymer can be considered as reliable as they fall within the applicability domain.

The structure–refractive index parallelism established by Eq. (4) succeeds in predicting the tendencies in data, i.e. high n values tend to be predicted as high and low n values as low. Now, although only one test set compound **225** is outside the applicability domain of Eq. (4), it is appreciated from the dispersion plot of residuals of Fig. 3 that four test set compounds have high residuals, and this can also be explained in terms of the data set heterogeneity: **231** (poly(p-xylylene)), **234** (poly(pentabromophenyl methacrylate)), **210** (poly(1,1-dichloroethylene)) and **183** (poly(2,3-dibromopropyl methacrylate)). In an attempt to improve these results, we considered the situation when these four polymers take part of the training set and the model was recalculated. This is justified, as the four compounds exhibit a higher residual than the rest of the test set compounds, and thus they should take part of the training set. The statistical quality found was somewhat better, with percentages of explained variances of 96%, 95% and 90% in train, val and test, respectively:

$$n = 0.0020 DCW_5 + 1.4759 \quad (5)$$

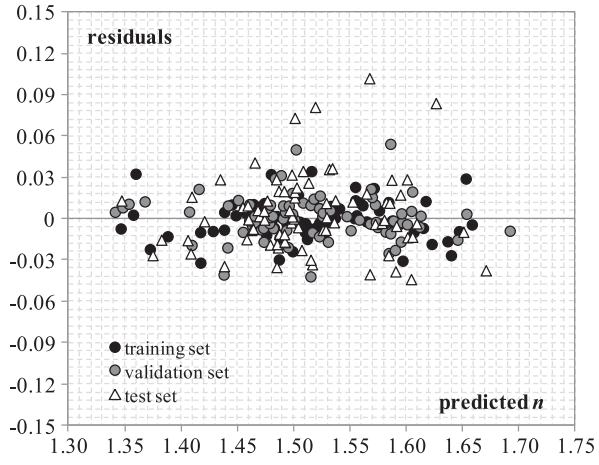

Fig. 3. Dispersion plot of residuals (Eq. (4)) for 234 polymer compounds.

$N_{train} = 82$, $R^2_{train} = 0.96$, $S_{train} = 0.0155$, $F = 1872$, $p<10^{-4}$, o(3S) = 1
$R^2_{loo} = 0.96$, $S_{loo} = 0.0159$, $R^2_{l20\%o} = 0.95$, $S_{l20\%o} = 0.0171$, $S^{rand} = 0.0649$
$N_{val} = 78$, $R^2_{val} = 0.95$, $S_{val} = 0.0158$
$N_{test} = 74$, $R^2_{test} = 0.90$, $S_{test} = 0.0218$.

In this equation, the parameters used for the DCW_5 calculation are $T = 1$ and $N_{epochs} = 8$. The QSPR of Eq. (5) represents an improvement over Eq. (4), as the ^2EC and ^3s_k structural attributes of the flexible descriptor are able to predict better the four aforementioned polymers having high residuals. The new model satisfies the necessary internal (Table 5S) and external validation conditions. Figs. 4 and 5 plot the predictions of Eq. (5) and its dispersion plot of residuals, respectively. There is only one outlier in the training set, **231** (poly(p-xylylene)), the same compound that appears previously posing a high residual in the test set. For the case of this new derived QSPR, all the compounds included in the test set belong to the applicability domain of the model (Table 6S). The correlation weights produced by the Monte Carlo simulation appear listed in Table 7S, while Table 8S includes an example for calculating DCW_5 for **1** (poly(pentadecafluorooctyl acrylate)).

From a total number of 249 structural attributes based on ^2EC and ^3s_k that can be obtained for the 234 polymers (Table 7S), only 190 of them contribute to the DCW_5 calculation. Furthermore, structural attributes with higher positive *CW* values, like EC2-S…9…, EC2-S…4…, 1… (… (…, S… (…C…, EC2-S…13…, O…=…C… or 2…c…1… tend to predict higher DCW_5 values and thus higher refractive indices. Here, EC2-Z…X… means that ^2EC_Z takes the value X, while A…B…C… is a three-element attribute. In addition, attributes with higher negative *CW* values as C…O…C…, Si…[…O…, EC2-N…13…, EC2-N…12…, EC2-Cl…6…, EC2-Si…19…, c…C… (… or C…N…C… lead to the opposite situation.

The ^2EC_Z index depends on the number of neighboring atoms to atom *Z*; when this index has zero-th order, then it equals the vertex degree in HSG. It is also known that the Morgan extended connectivity can be associated with the molecular symmetry. The higher the order of this index is, then the higher the atomic neighborhood features are seen by a considered atom. Finally, the ^3s_k index is a three-element SMILES attribute, revealing the importance of fragment elements in the prediction of *n* instead of individual SMILES elements.

The QSPR given by Eq. (5) predicts the refractive index of 234 structurally diverse polymers with a good accuracy, and compares favorably to previous published results. For instance, Eq. (4) analyzes a higher number of polymer compounds than other reported studies [20,22], and only involves a single descriptor when compared to the work of Bicerano [17] which uses 11 descriptors, or the work of García-Domenech and de Julián-Ortiz [18] having 10 descriptors. Furthermore, the developed model was properly internally and externally validated. Finally, another important result is that the calculated flexible descriptor does not require structural information on polymer molecular conformation, which means that the method is able to model the physical property by representing the molecular structure aspects with a similar or better degree of details as when using a 3D-geometry dependent approach [20,22].

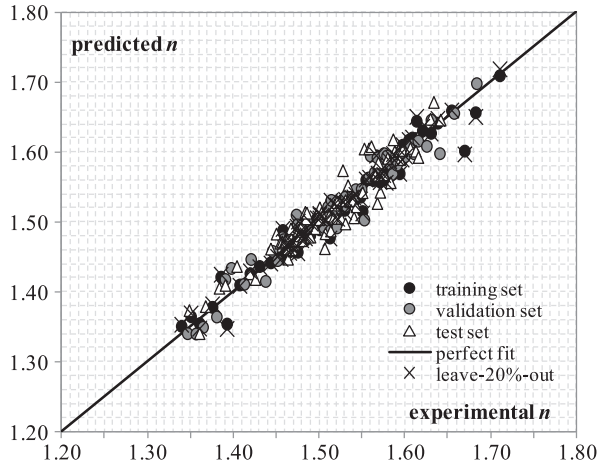

Fig. 4. Predicted (Eq. (5)) and experimental refractive indices for 234 polymer compounds.

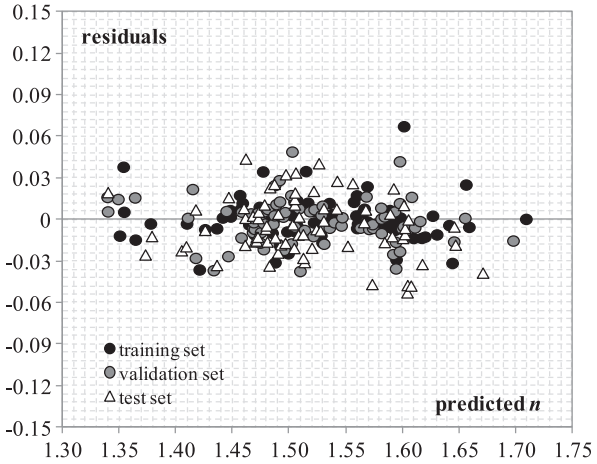

Fig. 5. Dispersion plot of residuals (Eq. (5)) for 234 polymer compounds.

4. Conclusions

The refractive index is considered a fundamental physical property of polymer compounds. In this work, we succeeded in proposing a polymer structure model that correlates the refractive index values with a good accuracy, and demonstrated that such model is predictive in the validation process. It is emphasized that the novelty of present work relies on the development of a structure–refractive index relationship for polymer macromolecules, through a computational technique that does not require the knowledge of the molecular conformation during the structural representation. The procedure employed here can be readily applied to the study of other polymer properties, which will be investigated in the near future.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgments

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. APT and AAT acknowledge support from EC project NANOPUZZLES (Project Reference: 309837). PRD, SEF and DEB are members of the scientific researcher career of CONICET.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.chemolab.2014.11.008.

References

[1] W. Knoll, Interfaces and thin films as seen by bound electromagnetic waves, Ann. Rev. Phys. Chem. 49 (1998) 569–638.

[2] R.J. Samuels, Application of refractive index measurements to polymer analysis, J. Appl. Polym. Sci. 26 (1981) 1383–1412.
[3] R. Hu, J.W.Y. Lam, B.Z. Tang, Recent progress in the development of new acetylenic polymers, Macromol. Chem. Phys. 214 (2013) 175–187.
[4] M. Schraub, S. Soll, N. Hampp, High refractive index coumarin-based photorefractive polysiloxanes, Eur. Polym. J. 49 (2013) 1714–1721.
[5] S.A. Sydlik, Z. Chen, T.M. Swager, Triptycene polyimides: soluble polymers with high thermal stability and low refractive indices, Macromolecules 44 (2011) 976–980.
[6] D.W. van Krevelen, K. te Nijenhuis, Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions, Elsevier, New York, 2009.
[7] C. Hansch, A. Leo, Exploring QSAR, Fundamentals and Applications in Chemistry and Biology, American Chemical Society, Washington, D. C, 1995.
[8] H. Kubinyi, QSAR: Hansch Analysis and Related Approaches, Wiley-Interscience, New York, 2008.
[9] T. Puzyn, J. Leszczynski, M.T.D. Cronin (Eds.), Recent Advances in QSAR Studies: Methods and Applications, Springer Science&Business Media B.V., Netherlands, 2010.
[10] A.R. Katritzky, E.V. Goordeva, Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research, J. Chem. Inf. Comput. Sci. 33 (1993) 835–857.
[11] M.V.E. Diudea, QSPR/QSAR Studies by Molecular Descriptors, Nova Science Publishers, New York, 2001.
[12] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics (Methods and Principles in Medicinal Chemistry), Wiley-VCH, Weinheim, 2009.
[13] A. Golbraikh, A. Tropsha, Beware of q 2! J. Mol. Graphics Modell. 20 (2002) 269–276.
[14] D.M. Hawkins, S.C. Basak, D.J. Mills, Assessing model fit by cross-validation, Chem. Inf. Model. 43 (2003) 579–586.
[15] M.V. Putz, A.-M. Putz, M. Lazea, L. Ienciu, A. Chiriac, Quantum-SAR extension of the spectral-SAR algorithm. Application to polyphenolic anticancer bioactivity, Int. J. Mol. Sci. 10 (2009) 1193–1214.
[16] M.V. Putz, C. Ionascu, A.-M. Putz, V. Ostafe, Alert-QSAR. Implications for electrophilic theory of chemical carcinogenesis, Int. J. Mol. Sci. 12 (2011) 5098–5134.
[17] J. Bicerano, Prediction of Polymer Properties, Marcel Dekker Inc., New York, 1996.
[18] R. García-Domenech, J.V. de Julián-Ortiz, Prediction of indices of refraction and glass transition temperatures of linear polymers by using graph theoretical indices, J. Phys. Chem. B 106 (2002) 1501–1507.
[19] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12.
[20] A.A. Katritzky, S. Sild, M. Karelson, Correlation and prediction of the refractive indices of polymers by QSPR, J. Chem. Inf. Comput. Sci. 38 (1998) 1171–1176.
[21] J. Xu, B. Chen, Q.J. Zhang, B. Guo, Prediction of refractive indices of linear polymers by a four-descriptor QSPR model, Polymer 45 (2004) 8651–8659.
[22] G. Astray, A. Cid, O. Moldes, J.A. Ferreiro-Lage, J.F. Gálvez, J.C. Mejuto, Prediction of refractive index of polymers using artificial neural networks, J. Chem. Eng. Data 55 (2010) 5388–5393.
[23] P.R. Duchowicz, N.C. Comelli, E.V. Ortiz, E.A. Castro, QSAR study for carcinogenicity in a large set of organic compounds, Curr. Drug Saf. 7 (2012) 282–288.
[24] A. Talevi, C.L. Bellera, M.D. Ianni, P.R. Duchowicz, L.E. Bruno-Blanch, E.A. Castro, An integrated drug development approach applying topological descriptors, Curr. Comput. Aided Drug Des. 8 (2012) 172–181.
[25] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, OCWLGI descriptors: theory and praxis, Curr. Comput. Aided Drug Des. 9 (2013) 226–232.
[26] L.M.A. Mullen, P.R. Duchowicz, E.A. Castro, QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents, Chemom. Intell. Lab. Syst. 107 (2011) 269–275.
[27] J. García, P.R. Duchowicz, M.F. Rozas, J.A. Caram, M.V. Mirífico, F.M. Fernández, E.A. Castro, A comparative QSAR on 1, 2, 5-thiadiazolidin-3-one 1, 1-dioxide compounds as selective inhibitors of human serine proteinases, J. Mol. Graphics Modell. 31 (2011) 10–19.
[28] E. Ibezim, P.R. Duchowicz, E.V. Ortiz, E.A. Castro, QSAR on aryl-piperazine derivatives with activity on malaria, Chemom. Intell. Lab. Syst. 110 (2012) 81–88.
[29] A.A. Toropov, A.P. Toropova, Prediction of heteroaromatic amine mutagenicity by means of correlation weighting of atomic orbital graphs of local invariants, J. Mol. Struct. Theochem 538 (2001) 287–293.
[30] A.A. Toropov, I.V. Nesterov, O.M. Nabiev, QSAR modeling of dihydrofolate reductase inhibitory activity by correlation weighting of nearest neighboring codes, J. Mol. Struct. Theochem 622 (2003) 269–273.
[31] A.A. Toropov, D. Leszczynska, J. Leszczynski, Predicting water solubility and octanol water partition coefficient for carbon nanotubes based on the chiral vector, J. Comput. Biol. Chem. 31 (2007) 127–128.
[32] A.A. Toropov, D. Leszczynska, J. Leszczynski, Predicting thermal conductivity of nanomaterials by correlation weighting technological attributes codes, Mater. Lett. 61 (2007) 4777–4780.
[33] A.A. Toropov, E. Benfenati, E. Additive, SMILES-based optimal descriptors in QSAR modelling bee toxicity: using rare SMILES attributes to define the applicability domain, Bioorg. Med. Chem. 26 (2008) 4801–4809.
[34] A.A. Toropov, A.P. Toropova, E. Benfenati, G. Gini, T. Puzyn, D. Leszczynska, J. Leszczynski, Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*, Chemosphere 89 (2012) 1098–1102.
[35] A.A. Toropov, A.P. Toropova, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, Comparison of SMILES and molecular graphs as the representation of the molecular structure for QSAR analysis for mutagenic potential of polyaromatic amines, Chemom. Intell. Lab. Syst. 109 (2011) 94–100.
[36] A.P. Toropova, A.A. Toropov, S.E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, CORAL: QSAR modeling of toxicity of organic chemicals towards *Daphnia magna*, Chemom. Intell. Lab. Syst. 110 (2012) 177–181.
[37] J. Brandrup, E.H. Immergut, E.A. Grulke, Polymer Handbook, 2nd ed. John Wiley and Sons, 1975.
[38] CORALSEA, http://www.insilico.eu/CORAL2011 (Accessed: 04-Aug-2014).
[39] E-Dragon, Milano chemometrics and QSAR research group. VCCLAB, virtual computational chemistry laboratory, http://michem.disat.unimib.it/chm (Accessed: 24-July-2014).
[40] CoDESSA, Katritzky group, http://www.ark.chem.ufl.edu/ (Accessed: 24-July-2014).
[41] C. Rücker, G. Rücker, M. Meringer, y-Randomization and its variants in QSPR/QSAR, J. Chem. Inf. Model. 47 (2007) 2345–2357.
[42] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.
[43] 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 (2003) 1361–1375.
[44] 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 (2011) 1575–1581.
[45] K. Roy, On some aspects of validation of predictive quantitative structure–activity relationship models, Expert Opin. Drug Discovery 2 (2007) 1567–1577.