Available online at www.sciencedirect.com
SCIENCE DIRECT
Polymer 47 (2006) 3240–3248
polymer
www.elsevier.com/locate/polymer

Prediction of intrinsic viscosity in polymer–solvent combinations using a QSPR model

Antreas Afantitis [a,b], Georgia Melagraki [a], Haralambos Sarimveis [a,*], Panayiotis A. Koutentis [c], John Markopoulos [d], Olga Igglessi-Markopoulou [a]

[a] School of Chemical Engineering, National Technical University of Athens, 9 Heroon Polytechniou St., 15780 Athens, Zografou, Greece
[b] Department of ChemoInformatics, NovaMechanics Ltd, Larnaca, Cyprus
[c] Department of Chemistry, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus
[d] Department of Chemistry, University of Athens, Athens, Greece

Received 23 December 2005; accepted 20 February 2006
Available online 20 March 2006

Abstract

In this work, a linear quantitative structure–property relationship (QSPR) model is presented for the prediction of intrinsic viscosity in polymer solutions. The model was produced by using the multiple linear regression (MLR) technique on a database that consists of 65 polymer–solvent combinations involving 10 different polymer. Among the 30 different physicochemical, topological and structural descriptors that were considered as inputs to the model, only eight variables (four variables for the polymer and four descriptors for the solvent) were selected using the elimination selection stepwise regression method (ES-SWR). The physical meaning of each descriptor is discussed in details. The accuracy of the proposed MLR model is illustrated using various evaluation techniques: leave-one-out cross validation procedure, validation through an external test set and *Y*-randomization. Furthermore, the calculation of the domain of applicability defines the area of reliable predictions.

Keywords: Intrinsic viscosity; QSPR; Molecular descriptors

1. Introduction

The design of new materials with optimal thermo physical, mechanical and optical properties is a challenge for computational chemistry. Novel materials are typically developed using a trial and error approach, which is costly and time-consuming [1]. An alternative strategy is to model the material properties as functions of the molecular structure using the so called quantitative structure–property relationships (QSPR) [2,3]. Application of QSPR methodologies in material design has the potential to decrease considerably the time and effort required to improve material properties in terms of their efficacy or to discover new materials with desired properties.

The conformational properties of polymer chains are usually determined in dilute solutions of the polymers. Viscosity, light scattering, small-angle X-ray scattering, and osmotic pressure, are the main types of measurement performed in solutions. According to Biceraco, conformational properties play a key role in determining the properties of polymer solutions, and therefore, in both synthesis (i.e. polymerization in solution) and processing (i.e. solvent casting of thin films) [4].

In particular, intrinsic viscosity (η) of polymer solutions, which we will investigate in the present work, is a measure of the volume associated with a given mass of polymer in a dilute, undisturbed solution at thermodynamic equilibrium [5].

Van Krevelen examined various equations for solution viscosity and recommended the following empirical equation (Eq. (1)) to estimate intrinsic viscosity when experimental measurements are not obtained under Θ conditions:

$$\eta \approx 0.99328\left(\frac{K}{M}\right)^2 \frac{\exp(8.5a^{10.3})M_v^{\alpha}}{M_{cr}^{(\alpha-0.5)}} \quad (1)$$

In the above equation M_{cr} is the critical molecular weight, M_v ($M_v = 2.5\times10^5$) is the viscosity average molecular weight, K is the molar stiffness function, M is the molecular weight per repeat unit, and the parameter α is defined by Eq. (2) that

* Corresponding author.
E-mail address: hsarimv@central.ntua.gr (H. Sarimveis).

0032-3861/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2006.02.060

follows next:

$$\begin{aligned} a &\approx 0.8 - 0.1\,\mathrm{abs}(\delta_{\mathrm{solvent}} - \delta_{\mathrm{polymer}}) \quad \text{for } \mathrm{abs}(\delta_{\mathrm{solvent}} - \delta_{\mathrm{polymer}}) \leq 3 \\ \alpha &\approx 0.5 \qquad\qquad\qquad\qquad\qquad\quad \text{for } \mathrm{abs}(\delta_{\mathrm{solvent}} - \delta_{\mathrm{polymer}}) > 3 \end{aligned} \tag{2}$$

where the solubility parameters (δ) are in $\sqrt{J/cc}$

Using Eq. (1) van Krevelen [4] calculated the intrinsic viscosity for 65 polymer–solvent combinations with a correlation coefficient $R^2 = 0.324$. The experimental values of the intrinsic viscosity are plotted against the values calculated by Eq. (1) in Fig. 1. After the rejection of seven solvents, which differ significantly from the polymer in hydrogen bonding capability, a better correlation coefficient ($R^2 = 0.483$) was obtained. The proposed equation works reasonably well unless the polymer has a tendency to be highly crystalline in the bulk. These last observations were also reported by Bicerano as weak points [4].

The limitations of empirical equations can be avoided with the use of QSPR approach. The physicochemical constants, quantum, topological and structural descriptors used in QSPR encode information about the structure of the module and thus implicitly account for cooperative effects between functional groups, charge redistribution and possible hydrogen bonding in the polymer [6,7]. The QSPR approach has been applied successfully to modeling many polymeric properties, such as glass transition temperature [2,8,9], refractive index [3,9,10] and solubility parameters [6].

In this work, we utilized the same series of 65 polymers–solvent combinations which involve 10 different polymers in order to determine the physicochemical and topological descriptors that correlate well with and can predict successfully the intrinsic viscosities in polymer solutions (in units of cm^3/g) [4]. The QSPR models were obtained by multiple linear regression (MLR). Thirty physicochemical and topological descriptors were calculated for each polymer and solvent using ChemSar which is included in the ChemOffice (CambridgeSoft Corporation) suite of programs [11]. Among them, the most statistically significant descriptors were selected, using the elimination selection stepwise regression (ES-SWR) variable selection method. The result of this study was the development of a new linear QSPR model containing eight variables. The proposed methodology was validated using several strategies: cross validation, *Y*-randomization and external validation using division of the available data set into training and test sets. Furthermore, the calculation of the domain of applicability defines the area of reliable predictions.

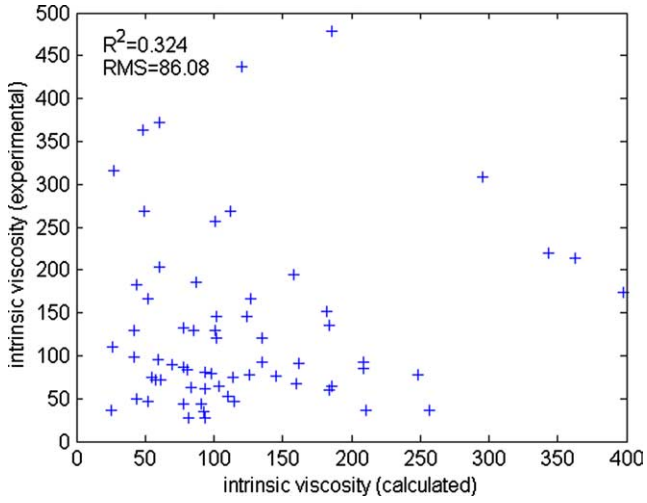

Fig. 1. Experimental versus calculated intrinsic viscosity using Eq. (1).

2. Materials and methods

2.1. Data set

For this QSPR study 65 polymer–solvent combinations together with their intrinsic viscosity were collected from Bicerano [4]. In order to model and predict this specific conformational property (intrinsic viscosity), 30 physicochemical, topological and structural descriptors for each polymer and solvent were considered as possible input candidates to the model (Table 1). All the descriptors were calculated using ChemSar (CambridgeSoft Corporation). The molecular descriptors were calculated from the structure of the monomer compound used in the polymerization.

2.2. Stepwise multiple regression

As mentioned in the introduction, the ES-SWR algorithm [12] was used to select the most appropriate descriptors. ES-SWR is a popular stepwise technique that combines forward

Table 1
Physicochemical constants, topological and structural descriptors

Id	Description	Notation
1	Molar refractivity	MR
2	Diameter	Diam
3	Partition coefficient (octanol water)	ClogP
4	Molecular topological index	TIndx
5	Principal moment of inertia *Z*	PMIZ
6	Number of rotatable bonds	NRBo
7	Principal moment of inertia *Y*	PMIY
8	Polar surface area	PSAr
9	Principal moment of inertia *X*	PMIX
10	Radius	Rad
11	Connolly accessible area	SAS
12	Shape attribute	ShpA
13	Connolly molecular area	MS
14	Shape coefficient	ShpC
15	Total energy	TotE
16	Sum of valence degrees	SVDe
17	Lumo energy	LUMO
18	Total connectivity	TCon
19	Humo energy	HUMO
20	Total valence connectivity	TVCon
21	Balaban index	BIndx
22	Wiener index	WIndx
23	Dipole length	DPLL
24	Electronic energy	ElcE
25	Repulsion energy	NRE
26	Connolly solvent-excluded volume	SEV
27	Ovality	Ovality
28	Cluster count	ClsC
29	Sum of degrees	SDeg
30	Molecular weight	MW

selection (FS-SWR) and backward elimination (BE-SWR). It is essentially a forward selection approach, but at each step it considers the possibility of deleting a variable as in the backward elimination approach, provided that the number of model variables is greater than two. The two basic elements of the ES-SWR method are described below in more details.

2.2.1. Forward selection

The variable considered for inclusion at any step is the one yielding the largest single degree of freedom F-ratio among the variables that are eligible for inclusion. The variable is included only if the corresponding F-ratio is larger than a fixed value F_{in}. Consequently, at each step, the jth variable is added to a k-size model if

$$F_j = \max_j \left(\frac{\text{RSS}_k - \text{RSS}_{k+j}}{s^2_{k+j}} \right) > F_{\text{in}} \tag{3}$$

In the above inequality RSS is the residual sum of squares and s is the mean square error. The subscript $k+j$ refers to quantities computed when the jth variable is added to the k variables that are already included in the model.

2.2.2. Backward elimination

The variable considered for elimination at any step is the one yielding the minimum single degree of freedom F-ratio among the variables that are included in the model. The variable is eliminated only if the corresponding F-ratio does not exceed a specified value F_{out}. Consequently, at each step, the jth variable is eliminated from the k-size model if

$$F_j = \min_j \left(\frac{\text{RSS}_{k-j} - \text{RSS}_k}{s^2_k} \right) < F_{\text{out}} \tag{4}$$

The subscript $k-j$ refers to quantities computed when the jth variable is eliminated from the k variables that have been included in the model so far.

2.3. Kennard and Stones algorithm

The Kennard and Stones algorithm [13] has gained an increasing popularity for splitting data sets into two subsets. The algorithm starts by finding two samples that are the farthest apart from each other on the basis of the input variables in terms of some metric, e.g. the Euclidean distance. These two samples are removed from the original data set and put into the calibration data set. This procedure is repeated until the desired number of samples has been reached in the calibration set. The advantages of this algorithm are that the calibration samples map the measured region of the input variable space completely with respect to the induced metric and that the test samples all fall inside the measured region. According to Tropsha [14] and Wu [15], Kennard and Stones algorithm is one of the best ways to build training and test sets.

2.4. Cross-validation technique

The reliability of the proposed method was explored using the cross-validation method. Based on this technique, a number of modified data sets are created by deleting in each case one or a small group (leave-some-out) of objects [16–18]. For each data set, an input–output model is developed, based on the utilized modelling technique. Each model is evaluated, by measuring its accuracy in predicting the responses of the remaining data (the ones that have not been utilized in the development of the model). In particular, the leave-one-out (LOO) procedure was utilized in this study, which produces a number of models, by deleting each time one object from the training set. Obviously, the number of models produced by the LOO procedure is equal to the number of available examples n. Prediction error sum of squares (PRESS) is a standard index to measure the accuracy of a modelling method based on the cross-validation technique. Based on the PRESS and SSY (sum of squares of deviations of the experimental values from their mean) statistics, the Q^2 and S_{PRESS} values can be easily calculated. The formulae used to calculate all the aforementioned statistics are presented below (Eqs. (5) and (6)):

$$Q^2 = 1 - \frac{\text{PRESS}}{\text{SSY}} = 1 - \frac{\sum_{i=1}^{n} (y_{\text{exp}} - y_{\text{pred}})^2}{\sum_{i=1}^{n} (y_{\text{exp}} - \bar{y})^2} \tag{5}$$

$$S_{\text{PRESS}} = \sqrt{\frac{\text{PRESS}}{n}} \tag{6}$$

For a more exhaustive testing of the predictive power of the model, except from the classical LOO cross-validation technique, validation of the model was carried out by a leave-five-out (L5O) cross validation procedure. From the training set we randomly selected groups of five compounds. Each group was left out and that group was predicted by the model developed from the remaining observations. This procedure was carried out several times, as will be shown in the sequel.

2.5. Y-randomization test

This technique ensures the robustness of a QSPR model [14,19]. The dependent variable vector (intrinsic viscosity) is randomly shuffled and a new QSPR model is developed using the original independent variable matrix. The new QSPR models (after several repetitions) are expected to have low R^2 and Q^2 values. If the opposite happens then an acceptable QSPR model cannot be obtained for the specific modeling method and data.

2.6. Estimation of the predictive ability of a QSPR model

According to Tropsha [14] the predictive power of a QSAR model can be conveniently estimated by an external $R^2_{\text{cv,ext}}$ (Eq. (7)).

$$R^2_{\text{cv,ext}} = 1 - \frac{\sum_{i=1}^{\text{test}} (y_{\text{exp}} - y_{\text{pred}})^2}{\sum_{i=1}^{\text{test}} (y_{\text{exp}} - \bar{y}_{\text{tr}})^2} \tag{7}$$

where $\bar{y}_{tr}$ is the averaged value for the dependent variable for the training set.

Furthermore, the same group [14,20] considered a QSAR model predictive, if the following conditions are satisfied:

$$R^2_{cv,ext} > 0.5 \tag{8}$$

$$R^2 > 0.6 \tag{9}$$

$$\frac{(R^2 - R^2_o)}{R^2} < 0.1 \text{ or } \frac{(R^2 - R'^2_o)}{R^2} < 0.1 \tag{10}$$

$$0.85 \le k \le 1.15 \text{ or } 0.85 \le k' \le 1.15 \tag{11}$$

Mathematical definitions of R^2_o, R'^2_o, k and k' are based on regression of the observed activities against predicted activities and vice versa (regression of the predicted activities against observed activities). The definitions are presented clearly in Golbraikh et al. [20] and are not repeated here for brevity.

2.7. Defining model applicability domain

The domain of application [14,21] of a QSPR model must be defined if the model is to be used for predicting properties of new combinations (polymer–solvent). Predictions for only those combinations that fall into this domain may be considered reliable. Extent of extrapolation [14] is one simple approach to define the applicability of the domain. It is based on the calculation of the leverage h_i [22] for each chemical, where the QSPR model is used to predict its property:

$$h_i = x_i^T (X^T X) x_i \tag{12}$$

In Eq. (12) x_i is the descriptor-row vector of the query compound and X is the $k \times n$ matrix containing the k descriptor values for each one of the n training compounds. A leverage value greater than $3k/n$ is considered large. It means that the predicted response is the result of a substantial extrapolation of the model and may not be reliable.

3. Results and discussion

For the selection of the most important descriptors, the aforementioned stepwise multiple regression technique was used. The procedure was automated by running a software package developed in our laboratory, which realizes the ES-SWR algorithm. The software was programmed in the Matlab programming language.

The descriptors that were selected using the ES-SWR algorithm were the following: HOMO, LUMO energies, principal moment of inertia X (PMIX) and molecular topological index (TIndx) for the solvents and dipole length (DPLL), molecular weight (MW), LUMO energy and Connolly molecular surface area (MS) for the polymers. Table 2 presents the correlation matrix, where it is clear that the eight selected descriptors are not highly correlated.

All the structures before the calculation of the descriptors were fully optimized using CS Mechanics and more specifically MM2 force fields and truncated-Newton–Raphson optimizer, which provide a balance between speed and accuracy [11].

A brief explanation of the descriptors that were selected follows next.

Molecular orbital (MO) surfaces visually represent the various stable electron distributions of a molecule. According to Frontier orbital theory, the shapes and symmetries of the highest-occupied and lowest-unoccupied molecular orbitals (HOMO and LUMO) are crucial in predicting the reactivity of a species and the stereochemical and regiochemical outcome of a chemical reaction [11]. Before calculating the HOMO and LUMO energies (eV) all the structures were additionally fully optimized using the AM1 basis set.

Molecular topological index (TIndx) [12] is derived from the adjacency matrix A, the distance matrix D and the A-dimensional column vector v, constituted by the vertex degree δ of the A atoms in the H-depleted molecular graph. The TIndx (also called Schultz index) is defined as:

$$\text{TIndx} = \sum_{i=1}^{A} [(A + D)v]_i = \sum_{i=1}^{A} t_i \tag{13}$$

where t_i are intricacy numbers of the A-dimensional column. Intricacy numbers measure the combined influence of valence, adjacency and distance for each comparable set of vertices; the lower the intricacy number, the more intricate or complex the vertex [12].

The principal moments of inertia (PMI) (g/mol Å^2) are physical quantities related to the rotational dynamics of a module [12]. The PMIs are defined by the diagonal elements of the inertia tensor matrix when the Cartesian coordinate axes are the principal axes of the module, with the origin located at the center of mass of the module. In this case the off-diagonal

Table 2
Correlation matrix of the eight selected descriptors

	HOMO(S)	LUMO(S)	TIndx(S)	PMIX(S)	DPLL(P)	MS(P)	LUMO(P)	MW(P)
HOMO(S)	1							
LUMO(S)	−0.040	1						
TIndx (S)	0.439	0.235	1					
PMIX (S)	−0.281	−0.276	0.203	1				
DPLL (P)	−0.052	−0.141	−0.367	−0.123	1			
MS (P)	−0.119	−0.133	−0.126	−0.105	0.353	1		
LUMO(P)	0.049	0.046	−0.184	0.096	0.303	−0.017	1	
MW (P)	−0.099	−0.183	0.090	−0.074	−0.223	−0.043	−0.717	1

Table 3
Calculated values for the 65 polymer–solvent combinations

Id	Polymer	Solvent	η (exp)	η (calc) $R^2=0.324$ RMS = 86.08 (Ref. [4]) (Fig. 1)	η (calc) $R^2=0.774$ RMS = 38.00 (Eq. (14)) (Fig. 2)	Leverages (limit 0.4154)
1	Polypropylene	Cyclohexane	295	309	207.32	0.1602
2		Toluene	182	151	181.90	0.1276
3		Benzene	160	120	141.22	0.2949
4	Polyisobutylene	Cyclohexane	209	186	113.11	0.2301
5		Carbon tetrachloride	135	95	51.52	0.1651
6		Toluene	87	74	63.25	0.1525
7		Benzene	59	63	40.53	0.4063
8	Polystyrene	Cyclohexane	42	49	87.69	0.1330
9		*n*-Butyl chloride	55	74	84.88	0.1137
10		Ethylbenzene	83	81	72.64	0.1891
11		Decalin	44	83	80.56	0.1297
12		Toluene	104	129	73.87	0.1100
13		Benzene	114	132	67.58	0.1212
14		Chloroform	94	182	73.04	0.0628
15		Butanone	52	195	89.54	0.0458
16		Chorobenzene	81	120	94.37	0.0295
17		Dioxane	85	79	106.97	0.0465
18	Poly(vinyl acetate)	Methyl isobutyl ketone	78	46	39.54	0.2164
19		Toluene	78	62	120.42	0.1506
20		3-Heptanone	44	72	108.01	0.1136
21		Benzene	94	76	163.50	0.2978
22		Chloroform	158	91	155.68	0.1133
23		Butanone	82	93	143.44	0.1633
24		Ethyl formate	102	145	144.67	0.1101
25		Chlorobenzene	98	166	157.26	0.1165
26		Dioxane	115	87	192.97	0.1347
27		Acetone	94	85	170.71	0.1844
28		Acetonitrile	101	36	50.57	0.0696
29		Methanol	61	36	89.74	0.1107
30	Poly(propylene oxide)	Toluene	145	166	37.47	0.1531
31		Benzene	162	316	45.08	0.0692
32	Poly(ethylene oxide)	Cyclohexane	186	89	86.93	0.0713
33		Carbon tetrachloride	135	145	74.69	0.1046
34		Benzene	120	269	80.09	0.0619
35		Chloroform	102	363	69.72	0.0680
36		Dioxane	127	219	106.27	0.2731
37		Acetone	78	214	43.51	0.0919
38		Dimethyl formamide	209	78	52.39	0.0982
39		Methanol	257	78	347.79	0.2637
40	Poly(methyl methacrylate)	Butyl chloride	25	35	280.15	0.3131
41		Methyl isoburyrate	42	36	124.91	0.2644
42		Methyl methacrylate	52	44	111.15	0.1364
43		Toluene	101	52	74.52	0.2169
44		Heptanone	27	60	139.11	0.0996
45		Ethyl acetate	60	65	136.31	0.0785
46		Benzene	70	68	179.09	0.1117
47		Chloroform	124	93	161.62	0.1618
48		Butanone	49	98	136.21	0.1001
49		Dichloroethane	60	65	133.40	0.0770
50		Tetrachloroethane	112	47	78.10	0.1275
51		Acetone	48	43	83.66	0.0412
52		Nitroethane	58	27	108.20	0.0755
53		Acetonitrile	26	27	105.40	0.0389
54	Polyacrylonitrile	Dimethyl acetamide	398	129	93.15	0.0806
55		Dimethyl formamide	343	479	98.56	0.0349
56		Dimethyl sulfoxide	363	437	110.81	0.1564
57		Butyrolactone	248	129	183.91	0.1910
58	Polybutadiene	Cyclohexane	126	257	53.57	0.0740
59		Isobutyl acetate	93	372	56.03	0.0644
60		Toluene	211	269	88.51	0.0766
61		Benzene	184	204	348.29	0.2637

Table 3 (*continued*)

Id	Polymer	Solvent	η (exp)	η (calc) $R^2=0.324$ RMS = 86.08 (Ref. [4]) (Fig. 1)	η (calc) $R^2=0.774$ RMS = 38.00 (Eq. (14)) (Fig. 2)	Leverages (limit 0.4154)
62	Polyisoprene	Hexane	91	72	344.73	0.2505
63		Isooctane	110	110	223.66	0.1931
64		Toluene	184	174	198.24	0.1400
65		Benzene	186	135	136.31	0.0785

elements of the inertia tensor matrix are zero and the three diagonal elements, I_{xx}, I_{yy}, and I_{zz} correspond to the moments of inertia about the *X*-, *Y*-, and *Z*-axis of the module. The ES-SWR algorithm identifies PMIX for the solvent as a significant descriptor for the modeling of intrinsic viscosity.

Dipole length (DPLL) is the electric dipole moment divided by the elementary charge. Electric dipole is a vector quantity, which encodes displacement with respect to the centre of gravity of positive and negative charges in a molecule [12].

The calculation of molecular surface area was made using Connolly's [23] method. Connolly molecular surface area (Å^2) is defined as the contact surface created when a spherical probe sphere (representing the solvent) is rolled over the molecular model. Molecular surface is a very important parameter of the molecules in understanding the structure and chemical behavior such as their ability to bind ligands and other molecules [12].

In order to investigate the possibility of having included outliers in our data set, the extent of the extrapolation method was applied to the 65 combinations that constitute the entire data set (Table 3). The leverages for all 65 compounds were computed (Table 3) and found to be inside the domain of the model (warning leverage limit 0.4154).

In the sequel we present the linear model, which was generated, with the eight most significant descriptors. The full linear equation for the prediction of the intrinsic viscosity (η) is the following:

$$\eta = 758 + 34.2\,\mathrm{HOMO(S)} + 14.6\,\mathrm{LUMO(S)} - 0.167\,\mathrm{TIndx(S)} + 0.388\,\mathrm{PMIX(S)} + 27.9\,\mathrm{DPLL(P)} - 0.364\,\mathrm{MS(P)} - 57.3\,\mathrm{LUMO(P)} - 3.18\,\mathrm{MW(P)} \quad (14)$$

$$\mathrm{RMS} = 38.00 \quad R^2 = 0.774 \quad F = 23.96 \quad Q^2 = 0.684$$

$$S_{\mathrm{Press}} = 44.95 \quad n = 65$$

Table 3 presents the experimental values of the intrinsic viscosity as well as the predictions using Eq. (1) and the newly proposed QSPR (Eq. (14)) for all the polymer–solvent combinations that constitute our data base. The experimental values of the intrinsic viscosity are also plotted against the values calculated by Eq. (14) in Fig. 2.

Molecules with high HOMO (highest occupied molecular orbital energy) values can donate their electrons more easily compared to molecules with low HOMO energy values, and hence are more reactive. Molecules with low LUMO (lowest unoccupied molecular orbital energy) values are more able to accept electrons than molecules with high LUMO energy values.

According to the above QSPR equation high HOMO and LUMO values for the solvent increase the intrinsic viscosity, however, the HOMO and LUMO coefficients are not equal. The larger HOMO coefficient indicates that the electron donating capability of the solvent molecule is more important than its electron accepting ability. Therefore, a solvent with the ability to donate electrons easily and accept electrons with difficulty should be used in order to increase the intrinsic viscosity. Similarly the intrinsic viscosity is increased when the monomer LUMO becomes increasingly negative. This correlates well with the high HOMO values predicted for the solvent molecules and supports the importance for strong interactions between solvent and polymer if a high intrinsic viscosity is desired.

Molecular topological index (TIndx) reduces the intrinsic viscosity due to a negative contribution to the QSPR equation. TIndx encodes solvent's information. The lower the TIndx value, the more intricate or complex is the molecule.

Principal moment of inertia along *x*- for the solvents gives information about how the product of mass and distance influence the value of intrinsic viscosity.

There is a strong relationship between polymers molecular weight and intrinsic viscosity [5]. The proposed QSPR model shows that as the monomer weight increases the intrinsic

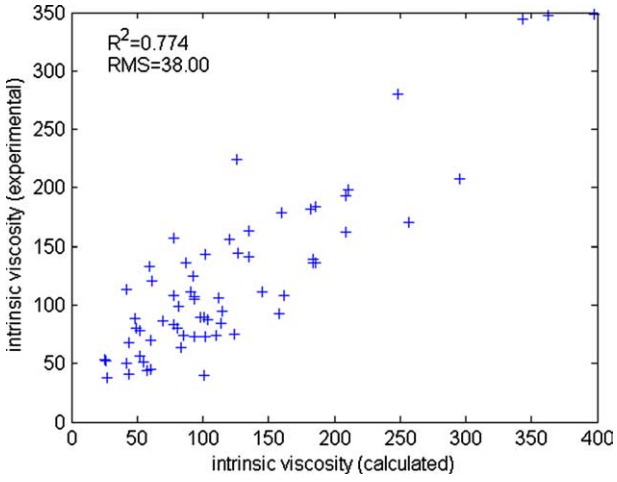

Fig. 2. Experimental versus calculated intrinsic viscosity using Eq. (14).

Table 4
Experimental and predicted values for the training and test set

Id	Polymer	Solvent	η (exp)	η (train) $R^2=0.759$ RMS = 34.67	η (pred) $R^2=0.751$ RMS = 49.39
1	Polypropylene	Cyclohexane	295	228.2150	
2		Toluene	182	204.4388	
3[a]		Benzene	160		201.94
4[a]	Polyisobutylene	Cyclohexane	209		181.95
5		Carbon tetrachloride	135	149.7865	
6[a]		Toluene	87		158.19
7[a]		Benzene	59		155.68
8	Polystyrene	Cyclohexane	42	118.3358	
9		*n*-Butyl chloride	55	58.2615	
10		Ethylbenzene	83	66.0278	
11		Decalin	44	29.5918	
12		Toluene	104	94.5672	
13		Benzene	114	92.0645	
14		Chloroform	94	73.7267	
15[a]		Butanone	52		86.67
16		Chorobenzene	81	87.1142	
17		Dioxane	85	77.7324	
18[a]	Poly(vinyl acetate)	Methyl isobutyl ketone	78		83.66
19[a]		Toluene	78		112.68
20		3-Heptanone	44	61.8575	
21[a]		Benzene	94		110.18
22[a]		Chloroform	158		91.84
23[a]		Butanone	82		104.78
24		Ethyl formate	102	78.5520	
25		Chlorobenzene	98	92.7184	
26		Dioxane	115	95.8449	
27		Acetone	94	114.5729	
28		Acetonitrile	101	41.7999	
29		Methanol	61	131.2033	
30[a]	Poly(propylene oxide)	Toluene	145		108.50
31		Benzene	162	105.9932	
32[a]	Poly(ethylene oxide)	Cyclohexane	186		178.68
33		Carbon tetrachloride	135	146.5188	
34		Benzene	120	152.4093	
35		Chloroform	102	134.0711	
36		Dioxane	127	138.0772	
37		Acetone	78	156.8052	
38		Dimethyl formamide	209	195.3810	
39		Methanol	257	173.4310	
40[a]	Poly(methyl methacrylate)	Butyl chloride	25		57.82
41		Methyl isoburyrate	42	48.4912	
42[a]		Methyl methacrylate	52		56.03
43		Toluene	101	94.1243	
44		Heptanone	27	31.0454	
45		Ethyl acetate	60	46.1533	
46		Benzene	70	91.6215	
47		Chloroform	124	73.2834	
48		Butanone	49	86.2296	
49		Dichloroethane	60	72.6592	
50		Tetrachloroethane	112	99.2702	
51[a]		Acetone	48		96.02
52		Nitroethane	58	46.4997	
53		Acetonitrile	26	59.2484	
54[a]	Polyacrylonitrile	Dimethyl acetamide	398		332.34
55[a]		Dimethyl formamide	343		331.85
56		Dimethyl sulfoxide	363	335.4329	
57		Butyrolactone	248	259.0258	
58[a]	Polybutadiene	Cyclohexane	126		245.92
59		Isobutyl acetate	93	137.5930	
60[a]		Toluene	211		222.15
61[a]		Benzene	184		159.99

Table 4 (*continued*)

Id	Polymer	Solvent	η (exp)	η (train) $R^2=0.759$ RMS = 34.67	η (pred) $R^2=0.751$ RMS = 49.39
62	Polyisoprene	Hexane	91	133.0560	
63		Isooctane	110	87.6832	
64		Toluene	184	162.4940	
65		Benzene	186	159.9911	

[a] The test set.

viscosity is reduced. This is expected since in order to maintain the viscosity average molecular weight M_v of 2.5×10^5 higher molecular weight monomers must give polymers with shorter chain lengths and thus lower intrinsic viscosities.

Molecular surface area (MS) encodes information for the polymers and explains their chemical behavior with other modules; high MS values reduce the viscosity.

The proposed QSPR model showed that solution intrinsic viscosity depends on polymer (MW) and solvent molecular weight (PMIX), polymer (MS) and solvent structure (PMIX, TIndx), the kind of interactions between polymer (MS, DPLL,) and solvents (PMIX) and finally the electronic behavior of polymer (DPLL, LUMO) and solvent (HOMO, LUMO) molecules.

The prediction ability of the selected descriptors was further explored using the data set of 65 polymer–solvent combinations which was divided into a training set of 45 combinations and a validation set of 20 combinations. The selection of the combinations in the training set was made according to the Kennard and Stones algorithm.

The combinations that constituted the training and validation sets are clearly presented in Table 4. The validation examples are marked with 'a'. The rest of the study will be concentrated on the model, which is constructed from the training set and will examine the predictive ability of the produced model. Using the same eight descriptors that were selected by the ES-SWR method, we developed a new MLR equation based on only the 45 training examples:

$$\eta = 824 + 38.4\,\mathrm{HOMO(S)} + 15.6\,\mathrm{LUMO(S)} - 0.188\,\mathrm{TIndx(S)} + 0.355\,\mathrm{PMIX(S)} + 15.0\,\mathrm{DPLL(P)} - 0.231\,\mathrm{MS(P)} - 56.2\,\mathrm{LUMO(P)} - 3.26\,\mathrm{MW(P)} \quad (15)$$

$$\mathrm{RMS} = 34.67 \quad R^2 = 0.759 \quad F = 14.15 \quad Q^2 = 0.601$$

$$S_{\mathrm{Press}} = 45.47 \quad n = 45$$

This equation was used to predict the intrinsic viscosity (η) (in units of cm^3/g) for the validation examples. The results are presented in the last column of Table 4 along with the respective R^2_{pred} statistic. The results illustrated once more that the linear MLR technique combined with a successful variable selection procedure is adequate to generate an efficient QSPR model for predicting the intrinsic viscosity (η) of different polymer–solvent combinations.

The proposed model (Eq. (15)) passed all the tests for the predictive ability (Eqs. (8)–(11))

$$R^2_{\mathrm{cv,ext}} = 0.749 > 0.5$$

$$R^2 = 0.759 > 0.6$$

$$\frac{(R^2 - R^2_{\mathrm{o}})}{R^2} = -0.662 < 0.1 \text{ or } \frac{(R^2 - R'^2_{\mathrm{o}})}{R^2} = -0.731 < 0.1$$

$$k = 0.9340 \text{ and } k' = 0.9834$$

It was important that the model was quite stable to the inclusion–exclusion of compound as measured by LOO and L5O correlation coefficients values, which are presented below:

$$Q^2 = 0.601$$

$$R^2_{\mathrm{cv\ L5O}} = 0.582$$

Calculation of the $R^2_{\mathrm{cv,L5O}}$ statistic was based on 1000 random selections of groups of five examples among the 45 training observations. Each group was left out and that group was predicted by the model developed from the remaining observations.

The model was further validated by applying the *Y*-randomization. Several random shuffles of the *Y* vector were performed and the results are shown in Table 5. The low R^2 and Q^2 values show that the good results in our original model are not due to a chance correlation or structural dependency of the training set.

It needs to be emphasized that no matter how robust, significant and validated a QSPR model may be, it cannot be expected to reliably predict the modeled property for the entire universe of chemicals. The extrapolation method was applied

Table 5
R^2 and Q^2 values after several *Y*-randomization tests

Iteration	R^2	Q^2
1	0.15	0.00
2	0.16	0.00
3	0.05	0.00
4	0.32	0.10
5	0.08	0.00
6	0.28	0.00
7	0.19	0.00
8	0.09	0.00
9	0.11	0.00
10	0.29	0.09

Table 6
Leverages for the test set

Combination Id	Leverages
3	0.2092
4	0.2855
6	0.1950
7	0.1532
15	0.1522
18	0.0568
19	0.1161
21	0.0607
22	0.0936
23	0.0492
30	0.2367
32	0.3058
40	0.0943
42	0.0911
51	0.1035
54	0.5874
55	0.5528
58	0.3767
60	0.2779
61	0.1579

to the compounds that constitute the test set. The leverages for all 20 compounds were computed (Table 6). Not one of the 20 compounds fell outside from the domain of the model (warning leverage limit 0.60).

The proposed method, due to the high predictive ability [14,24], can be a useful aid to the costly and time consuming experiments for determining the intrinsic viscosity. The method can also be used to screen existing databases or virtual combinations in order to identify combinations with desired intrinsic viscosity. In this case, the applicability domain will serve as a valuable tool to filter out 'dissimilar' combinations.

4. Conclusion

A novel QSPR model was developed in this work that can predict intrinsic viscosity using molecular descriptors. Using a data set of 65 polymer–solvent combinations and a rigorous variable selection method, eight descriptors were chosen among the 30 different descriptors that were examined. Several validation techniques illustrated the accuracy of the produced model not only by calculating its fitness on sets of training data but also by testing the predicting abilities of the model. The encouraging results showed that the QSPR methodology can be used successfully for deriving estimators for the intrinsic viscosity and other polymer properties. Moreover, it overcomes several of the limitations experienced by empirical models.

Acknowledgements

A.A. wishes to thank Cyprus Research Promotion Foundation (grant no. PENEK/ENISX/0603/05) and A.G. Leventis Foundation for financial support. G.M. thanks the Greek State Scholarship Foundation for a doctoral assistantship. The authors thank Costas Patrickios for helpful discussions.

References

[1] Camarda KV, Maranas CD. Ind Eng Chem Res 1999;38:1884–92.
[2] Katritzky AR, Sild S, Lobanov V, Karelson M. J Chem Inf Comput Sci 1998;38:300–4.
[3] Katritzky AR, Sild S, Karelson M. J Chem Inf Comput Sci 1998;38:1171–6.
[4] Bicerano J. Prediction of polymer properties. 2nd ed. New York: Marcel Dekker; 1996.
[5] Rushing TS, Hester RD. Polymer 2004;45:6587–94.
[6] Yu X, Wang X, Wang H, Li X, Gao J. QSAR Comb Sci 2005;2:151–61.
[7] Yu X, Wang X, Gao J, Li X, Wang H. Polymer 2005;46:943–9451.
[8] Afantitis A, Melagraki G, Makridima K, Alexandridis A, Sarimveis H, Igglessi-Markopoulou O. J Mol Str: TheoChem 2005;716:193–8.
[9] Domenech AG, de Julian-Ortiz JV. J Phys Chem B 2003;106:1501–7.
[10] Xu J, Chen B, Zhang Q, Guo B. Polymer 2004;45:8651–9.
[11] CambridgeSoft Corporation (www.cambridgesoft.com).
[12] Todeschini R, Consonni V, Mannhold R, (Series Editor), Kubinyi H, (Series Editor), Timmerman H, (Series Editor) Handbook of molecular descriptors. Weinheim: Wiley–VCH; 2000.
[13] Kennard RW, Stone LA. Technometrics 1969;11:137–48.
[14] Tropsha A, Gramatica P, Gombar VK. Quant Struct Activ Relat 2003;22:1–9.
[15] Wu W, Walczak B, Massart DL, Heuerding S, Erni F, Last IR, et al. Chemom Intell Lab Syst 1996;33:35–46.
[16] Efron B. J Am Stat Assoc 1983;78:316–31.
[17] Efroymson MA. Multiple regression analysis. In: Ralston A, Wilf HS, editors. Mathematical methods for digital computers. New York: Wiley; 1960.
[18] Osten DW. J Chemom 1998;2:39–48.
[19] Wold S, Eriksson L. Statistical validation of QSAR results. In: van de Waterbeemd H, editor. Chemometrics methods in molecular design. Weinheim (Germany): Wiley–VCH; 1995. p. 309–18.
[20] Golbraikh A, Tropsha A. J Mol Graphics Modell 2002;20:269–76.
[21] Shen M, Beguin C, Golbraikh A, Stables J, Kohn H, Tropsha A. J Med Chem 2004;47:2356–64.
[22] Atkinson A. Plots, transformations and regression. Oxford (UK): Clarendon; 1985 p. 282.
[23] Connolly M. J Mol Graph 1993;11:139–41.
[24] Aptula AO, Jeliazkova NG, Schultz TW, Cronin MTD. QSAR Comb Sci 2005;24:385–96.