Green Chemistry

RSC Publishing

PAPER

View Article Online
View Journal | View Issue

Cite this: *Green Chem.*, 2013, **15**, 2283

Quantitative structure–property relationships prediction of some physico-chemical properties of glycerol based solvents†

José I. García,*[a] Héctor García-Marín,[a] José A. Mayoral[a,b] and Pascual Pérez[c]

Received 12th April 2013,
Accepted 12th June 2013

DOI: 10.1039/c3gc40694f

www.rsc.org/greenchem

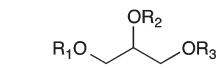
Quantitative structure–properties relationships (QSPR) models have been developed for three characteristic properties of a series of 62 new glycerol derivatives, relevant to solvent classification and substitution uses. Using structural descriptor variables, three equations have been found using multiple linear regression analysis, which can be applied for *in silico* prediction of physico-chemical properties, allowing a faster selection of target solvents for a given application.

Introduction

Organic solvents are used in huge amounts in many industrial and daily life applications, but unfortunately the majority of them come from petroleum and they are often labelled as toxic or hazardous substances. For this reason, substantial efforts are being made to develop more benign solvents from renewable sources. Our group has recently described a family of solvents based on glycerol,[1] a concomitant product in biodiesel production. To facilitate the search for possible substitution applications, we have also determined a number of physico-chemical properties of these glycerol-derived solvents, and compared them with those of conventional organic solvents. Many of these properties are difficult to measure, so it is clear that the development of efficient quantitative structure–properties relationship (QSPR) equations would be of great interest to accelerate the search for the best solvent for a given application. The concept is based on the fact that there exists a close relationship between the bulk properties and the molecular structure of a series of similar chemical compounds.

In this context, solvent classification is a very interesting issue, which has traditionally been addressed from both microscopic (intermolecular interactions) and macroscopic (as a continuum medium) approaches. However, solvation processes are hard to parameterize given that solvation energy (the only observable magnitude) is controlled by a large number of factors. For this reason, classification of solvents, and especially that of neoteric solvents, is far from being straightforward,[2–7] and hence, during the last three decades of the 20th century, many efforts have been devoted to classifying solvents using empirical parameters.[8]

Quantitative structure–property relationships are mathematical equations relating chemical structure to a wide variety of physical, chemical, and biological properties; in our case, solvent properties. QSPR models, once established, can be used to predict the properties of compounds as yet unmeasured or even unknown.[9–13] In this context, there are many reports about the applications of QSPR in connection with solvents, such as physico-chemical properties in alkane series,[14] optical properties of organic compounds,[15] thermophysical properties of some fluids,[16] solubility of hazardous compounds,[17] acidity constants of some acid derivatives,[13] permeability of organic compounds in membranes,[18] or important properties of room temperature ionic liquids (RTILs), such as toxicity.[19–21] A major step in the development



R =	Code =
H	0
Me	1
Et	2
^{i}Pr	3i
^{n}Bu	4
^{i}Bu	4i
^{t}Bu	4t
CF_3CH_2	3F
$CF_3CF_2CH_2$	5F
$CF_3(CF_2)_2CH_2$	7F

Scheme 1 General structure and codification of the glycerol-derived solvents used in this study.

[a]*Instituto de Síntesis Química y Catálisis Homogénea, Facultad de Ciencias, CSIC-Univ. de Zaragoza, Pedro Cerbuna, 12, E-50009 Zaragoza, Spain. E-mail: jig@unizar.es; Tel: +34 976762271*
[b]*Department of Organic Chemistry, Facultad de Ciencias, Univ. de Zaragoza, Pedro Cerbuna, 12, E-50009 Zaragoza, Spain*
[c]*Department of Physical Chemistry, Facultad de Ciencias, Univ. de Zaragoza, Pedro Cerbuna, 12, E-50009 Zaragoza, Spain*

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c3gc40694f

Table 1 List of properties of the 62 glycerol solvents studied in the present work

Solvent	Code	E_T^N	η (cP)	b.p. (°C)
1,2,3-Propanetriol	000	0.812	934[30]	290[30]
3-Methoxy-1,2-propanediol	100	0.710	37.72	222
3-Ethoxy-1,2-propanediol	200	0.690	35.14	221
3-*n*-Butoxy-1,2-propanediol	400	0.680	42.03	249
1,3-Dimethoxy-2-propanol	101	0.610	3.46	170
1-Isopropoxy-3-methoxy-2-propanol	103i	0.490	3.38	188
1-*n*-Butoxy-3-methoxy-2-propanol	104	0.480		208
1-Isobutoxy-3-methoxy-2-propanol	104i	0.490		200
1-*tert*-Butoxy-3-methoxy-2-propanol	104t	0.440		195
1-*n*-Butoxy-3-isopropoxy-2-propanol	403i	0.470	4.59	223
1,3-Di-*n*-butoxy-2-propanol	404	0.450	5.53	248
3-*n*-Butoxy-1-*tert*-butoxy-2-propanol	404t	0.390		230
3-*n*-Butoxy-1-isobutoxy-2-propanol	404i	0.460		229
1-Ethoxy-3-isopropoxy-2-propanol	203i	0.450		187
1-*n*-Butoxy-3-ethoxy-2-propanol	204	0.450		220
1-*tert*-Butoxy-3-ethoxy-2-propanol	204t	0.410		204
1-Isobutoxy-3-ethoxy-2-propanol	204i	0.460		214
1,3-Diisopropoxy-2-propanol	3i03i	0.460		202
1-*tert*-Butoxy-3-isopropoxy-2-propanol	3i04t	0.370		202
1-Isobutoxy-3-isopropoxy-2-propanol	3i04i	0.440		215
1-Isopropoxy-3-(2,2,2-trifluoroethoxy)-2-propanol	3i03F	0.590	6.90	176
1-*n*-Butoxy-3-(2,2,2-trifluoroethoxy)-2-propanol	403F	0.600		210
1-*tert*-Butoxy-3-(2,2,2-trifluoroethoxy)-2-propanol	4t03F	0.570	8.61	199
1-Isobutoxy-3-(2,2,2-trifluoroethoxy)-2-propanol	4i03F	0.600		205
1,3-Bis(2,2,2-trifluoroethoxy)-2-propanol	3F03F	0.700	8.14	197
1,3-Bis(2,2,3,3,3-pentafluoropropoxy)-2-propanol	5F05F	0.699		204
1,3-Bis(2,2,3,3,4,4,4-heptafluorobutoxy)-2-propanol	7F07F	0.685	19.60	206
1,2,3-Trimethoxypropane	111			150
1-Isopropoxy-2,3-dimethoxypropane	113i		1.03	170
2-*n*-Butoxy-3-methoxy-1-isopropoxypropane	143i	0.145		215
1-*tert*-Butoxy–2,3-dimethoxypropane	114t	0.214		180
2-*n*-Butoxy-1-*tert*-butoxy-3-methoxypropane	144t			219
1-*n*-Butoxy-2,3-dimethoxypropane	114	0.178		199
1,2-Di-*n*-butoxy-3-methoxypropane	144			234
1-Isobutoxy-2,3-dimethoxypropane	114i			193
2-*n*-Butoxy-1-isobutoxy-3-methoxypropane	144i			227
2-Ethoxy-3-methoxy-1-isopropoxypropane	123i	0.167		161
3-Ethoxy-2-methoxy-1-isopropoxypropane	213i	0.171		183
1-*tert*-Butoxy-2-ethoxy-3-methoxypropane	124t			190
1-*tert*-Butoxy-3-ethoxy-2-methoxypropane	214t	0.150		193
1-*n*-Butoxy-2-ethoxy-3-methoxypropane	124	0.155		209
1-*n*-Butoxy-3-ethoxy-2-methoxypropane	214	0.164		209
1-Isobutoxy-2-ethoxy-3-methoxypropane	124i			198
1-Isobutoxy-3-ethoxy-2-methoxypropane	214i	0.161		201
2,3-Diethoxy-1-isopropoxypropane	223i			192
1-*tert*-Butoxy-2,3-diethoxypropane	224t	0.155		199
1-*n*-Butoxy-2,3-diethoxypropane	224	0.161		217
1-Isobutoxy-2,3-diethoxypropane	224i			210
1-*n*-Butoxy-2-methoxy-3-isopropoxypropane	413i	0.155	1.67	218
1-*n*-Butoxy-2-ethoxy-3-isopropoxypropane	423i			222
3-*n*-Butoxy-1-*tert*-butoxy-2-methoxypropane	414t	0.141		234
3-*n*-Butoxy-1-*tert*-butoxy-2-ethoxypropane	424t			211
1,3-Di-*n*-butoxy-2-methoxypropane	414	0.145	3.78	244
3-*n*-Butoxy-1-isobutoxy-2-methoxypropane	414i	0.150		226
3-*n*-Butoxy-1-isobutoxy-2-ethoxypropane	424i			241
3-Isopropoxy-2-methoxy-1-(2,2,2-trifluoroethoxy)-propane	3i13F			180
3-*tert*-Butoxy-2-methoxy-1-(2,2,2-trifluoroethoxy)-propane	4t13F	0.373	2.14	185
1,2,3-Tri-*n*-butoxypropane	444		2.72	270
3-*n*-Butoxy-2-methoxy-1-(2,2,2-trifluoroethoxy)propane	413F			207
2-Methoxy-1,3-bis(2,2,2-trifluoroethoxy)propane	3F13F	0.553	2.33	178
2-Ethoxy-1,3-bis(2,2,2-trifluoroethoxy)propane	3F23F	0.595		171
2-*n*-Butoxy-1,3-bis(2,2,2-trifluoroethoxy)propane	3F43F	0.574		208

of QSPR models is finding a set of molecular descriptors able to represent the variation of the structural features of the molecules, and therefore a wide variety of descriptors have been reported for use in QSPR analysis.[22–25] The molecular descriptors chosen (X) are correlated with one or more response variables (Y) using different statistical approaches. Among the many statistical procedures available to establish those relationships, such as Partial Least Squares Analysis

(PLSA), Multiple Linear Regression (MLR), Artificial Neural Network (ANN), or Principal Component Analysis (PCA), a really good example of using QSPR in the classification of solvents through PCA can be found in the literature.[26] Probably MLR[27] is the most widely used because it is simple and intuitive.

From an industrial application point of view, there are three main solvent features that must be taken into account: (1) the behaviour of dissolution processes, which can be well defined through the solvatochromic parameter E_T^N (see below),[28,29] (2) mechanical aspects, which can be quantified by their viscosity, and (3) volatility aspects, closely related to safety, toxicity and air pollution, which can be considered through the boiling point.

Therefore, we decided to select for the present work 62 solvents based on glycerol, all of them prepared in our laboratory (Scheme 1 and Table 1),[1] and the three above-mentioned properties, also determined by us, were analyzed for this solvent set using several QSPR models.

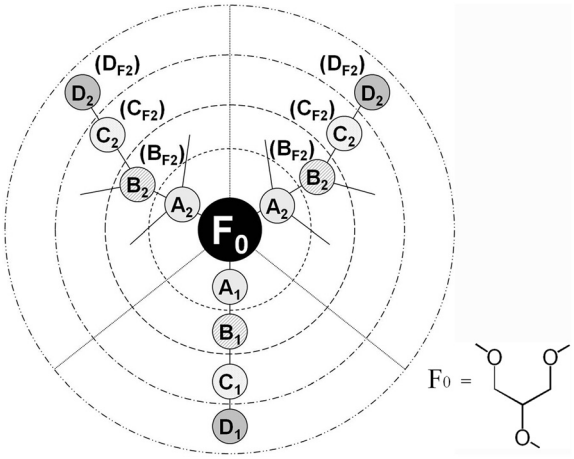


Fig. 1 The DARC/PELCO scheme used to describe glycerol based solvent structures.

Results and discussion

Molecular structure definition

There are many ways of describing the structure of a chemical compound as a vector of numbers. In this work we have used two different approaches, based on molecular connectivity descriptors: topological parameters and DARC/PELCO descriptors.

Topological parameters are based on the molecular graph of each compound.[25,31] They are easily determined from the connectivity and adjacency matrixes of each compound. The number of connected components of a graph is a topological invariant that measures the number of structurally independent or disjoint subnetworks. These parameters are excellent descriptors of molecular size, shape and flexibility. They are global parameters in the sense that the whole molecular structure is condensed in a single number. The topological descriptors selected for QSPR studies in this work are: (i) hydrogen bond acceptor counters (HBA), (ii) hydrogen bond donor counters (HBD), (iii) rotatable bond counters (RB), (iv) flexibility index (φ),[32] (v) Balaban index (Bal),[33] (vi) Wiener index (W),[34] (vii) Zagreb index (Z),[35] (viii) Kier shape index (κ_n),[32] (ix) subcount index (SC),[36] and (x) connectivity index (χ).[25] Full definition of the indices used in the statistical analyses are given in the ESI.†

DARC/PELCO (Description, Acquisition, Retrieval and Computer-aided design/Perturbation of an Environment which is Limited, Concentric and Ordered)[37] is another excellent way to describe chemical structures, yet is much less used in QSPR studies. This system is particularly suitable for studying families of compounds with a common chemical substructure. The DARC/PELCO method is based on the exhaustive generation of all topochromatic sites around the reference structure (F_0), which corresponds to the glycerol skeleton common to all structures, and the evaluation of their contribution to the property. The DARC/PELCO descriptors are local, since each one indicates the presence or absence of a group of atoms in a given molecular position. Their definition is shown in Fig. 1. In this definition we have incorporated the symmetry of the glycerol derivatives used, by assuming that the contributions of groups occupying equivalent positions (*i.e.* those linked to carbons 1 and 3 of the glycerol moiety) will display the same influence on the property under study. Preliminary studies have demonstrated that this simplification does not alter the results of the regression analyses.

Solvent properties selection

Solvent polarity (E_T^N). Solvent polarity parameters have demonstrated their usefulness not only to classify organic solvents but also to explain solvent effects on very different physical and chemical processes. An excellent overview of solvent polarity parameters and their applications can be found in Reichardt's outstanding book.[8] Although there are several procedures to quantify solvent polarity, solvatochromism measurements of probe dyes are undoubtedly the most successful methodology for an accurate determination of this solvent feature due to their easy determination and their high sensitivity to small polarity changes. From this point of view, the Dimroth and Reichardt $E_T(30)$ parameter[28,29] is one of the most widely used parameters. $E_T(30)$ values represent a blend of dipolarity/polarizability and hydrogen bond donor solvation abilities of the solvent, the latter feature contributing to the total $E_T(30)$ value to a greater extent. E_T^N is a normalized form of $E_T(30)$, taking the value 0 for hexadecane and 1 for water.

Viscosity. Viscosity describes a fluid's internal resistance to flow and may be thought of as a measure of fluid friction. This property is particularly interesting from the viewpoint of possible large scale industrial applications, where big solvent volumes have to be stirred and pumped from one place to another.

Boiling point. One major problem concerning the use of organic solvents is the presence of traces of these compounds in the air. The most common volatile organic compounds (VOC's) are solvents indeed. Nowadays a great deal of effort is

being made to solve this problem, trying to substitute these volatile solvents with others that are less or not volatile. For this reason, this property is really important to be not only measured but also predicted. Boiling point is a quick and easy form to estimate the volatility of a solvent, since in general a higher boiling point correlated with a lower volatility at ambient pressure and temperature.

Quantitative structure–properties relationships

Multiple linear regression (MLR) with topological indices. It is often assumed that the relationship between structural parameters and experimental properties is well represented by a linear model:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \text{ or } Y = X \cdot B \text{ (in matrix form)} \quad (1)$$

In eqn (1) the b_i are unknown coefficients, and the objective of regression analysis is to estimate their values. As QSPR data sets consist of variables that are diverse in range, variation and size, prior to regression analysis auto-scaling is usually applied, *i.e.*, the *i*th column is mean centred (with x_i) and scaled with $1/\mathrm{SD}(x_i)$, where SD is the standard deviation. When X is of full rank the least squares solution is: $\boldsymbol{B} = (\boldsymbol{X}^{\mathrm{T}}\cdot\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$, where $\boldsymbol{B}$ is the estimator vector for the regression coefficients. However, very often, not all these coefficients have statistical significance, so the final QSPR model should only keep those descriptors really contributing to the variation in the property observed. To this end we used a stepwise method for variable selection. In this way, independent variables x_i are entering and leaving in the regression equation, and only those having statistically significant coefficients are finally kept in the model fitting.

The three regression equations obtained for the three experimental properties fitted are the following:

$$E_{\mathrm{T}}^{\mathrm{N}} = b_0 + b_1 \cdot \mathrm{HBA} + b_2 \cdot \mathrm{HBD} + b_3 \cdot \mathrm{SC}_3^{\mathrm{p}} \quad (2)$$

$$\eta = b_0 + b_1 \cdot \mathrm{HBD} \quad (3)$$

$$\mathrm{bp} = b_0 + b_1 \cdot \mathrm{RB} + b_2 \cdot \mathrm{HBD} + b_3 \cdot \mathrm{HBA} \quad (4)$$

The corresponding coefficient values and MLR parameters are shown in Table 2.

As can be seen, the hydrogen-bonding ability of the solvent seems to be the most important feature in modeling the three properties under study. This result is consistent with the kind of intermolecular interactions involved. It is well-known that $E_{\mathrm{T}}^{\mathrm{N}}$ values are dominated by the HBD ability of the solvent, due to the strong specific solvation established through hydrogen-bonding with the phenolate oxygen of the betaine dye. Similarly, the strong solvent–solvent intermolecular hydrogen-bond interactions of most of the glycerol-derived solvents included in the study are at the origin of the viscosity values obtained, and hence the importance of this coefficient in the MLR model. Finally, the same strong intermolecular interactions can be invoked to explain the high boiling points displayed by most of the solvents considered.

Table 2 Linear regression parameters from eqn (2)–(4)[a]

	$E_{\mathrm{T}}^{\mathrm{N}}$	η	b.p.
$b_0 \pm e_0$	0.206 ± 0.035	—[b]	111.1 ± 17.0
$b_1 \pm e_1$	0.073 ± 0.010	14.50 ± 3.56	11.8 ± 1.9
$b_2 \pm e_2$	0.194 ± 0.021	—	24.7 ± 5.0
$b_3 \pm e_3$	−0.019 ± 0.004	—	−3.2 ± 1.2
N	46	17	62
R^2	0.957	0.823	0.769
$\sigma(y)$	0.0437	7.51	12.2
F[c]	72.39	74.57	64.52

[a] b_i are the coefficients for each regression, e_i is the tolerance for the b_i value in a 95% confidence interval. N is the number of cases (solvents data) used in each regression, R^2 is the determination coefficient. [b] As the b_0 coefficient turned out to be non-significant in the standard MLR analysis, fitting was done by forcing the equation to pass through the origin of coordinates. A slight improvement in R^2 was obtained in this way. [c] $F_{(3,\ 42,\ 0.05)} = 2.84$; $F_{(1,\ 16,\ 0.05)} = 4.41$; $F_{(3,\ 58,\ 0.05)} = 2.84$. All equations are statistically significant ($p > 95\%$).

Fig. 2 plots the experimental values *vs.* those calculated with the three MLR models. The dotted line represents the least squares fit between both sets of data.

As can be seen the best results are obtained in the case of the $E_{\mathrm{T}}^{\mathrm{N}}$ solvation parameter, which is consistent with the higher determination coefficient value obtained in the MLR analysis. In the other two cases, although there is a clear correlation, as indicated by the grouping of points around the diagonal, the fit is not good enough to lead to a fully predictive model.

The robustness and predictivity character of the method were tested by splitting the data into a training and a test set, which was created by extracting eight solvents from the complete set, so the training set consists of 54 solvents. The solvents of the test set (Table 3) were selected bearing in mind the representativity of the whole set, and for all the properties the test set size is within the usually recommended percentage of 10–20% of the total cases.

The three new regression equations obtained with the new 54 solvents group of the training set are summarized in Table 4.

As can be seen, the regression coefficients are in all cases very close to those calculated with the full set of solvents, which illustrates the robustness of the equations obtained. These new equations were used to predict the polarity, viscosity and boiling point of solvents in the test set. As a measure of the goodness of the prediction we used the mean unsigned error (MUE). In the case of $E_{\mathrm{T}}^{\mathrm{N}}$ the MUE of the fitting of the training set was 0.028, whereas that of the predictions of the test set was 0.030 and represents less than 5% of the whole range of values (0.671). This points to a reasonable predictivity for the model developed. In the case of viscosity the corresponding MUE values for the training and test sets are 7.28 and 5.08, respectively, *i.e.* 18% of the whole range of values (41.0) in the worst case, which indicates the poorer predictivity of the corresponding equations, although they could still be used in a semi-quantitative way. Finally, in the case of the boiling points, the MUE values for the training and test sets are 8.6

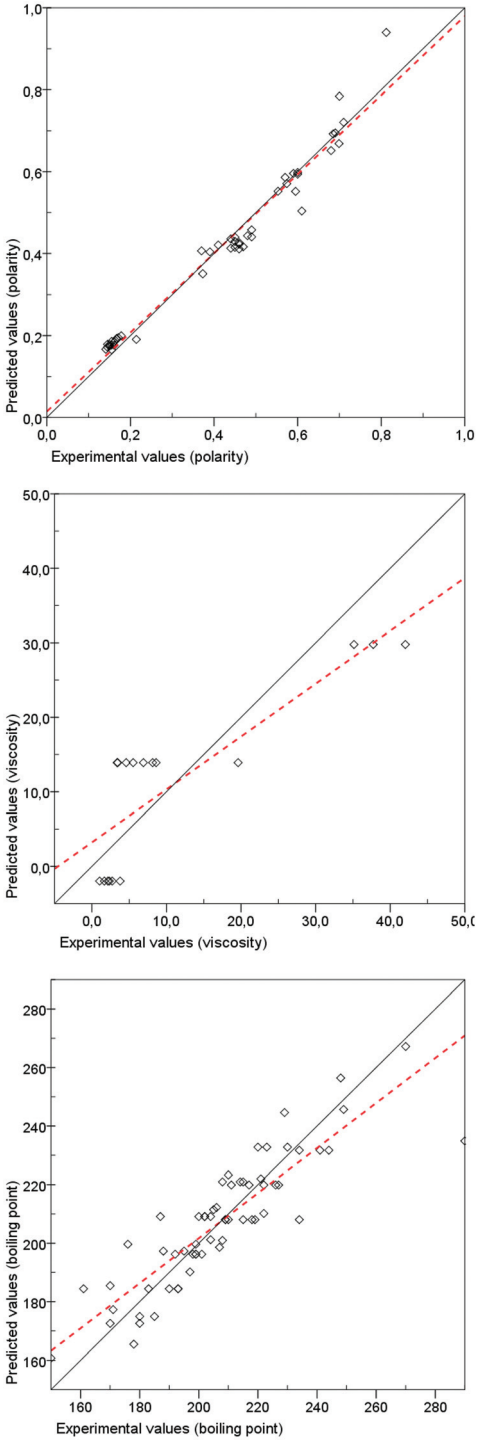


Fig. 2 Plots of predicted *vs.* experimental values of E_T^N (a), viscosity (b), and boiling point (c), as calculated through MLR analysis using topological indices.

and 10.9, respectively. Again, the error is only slightly higher in the case of the "pure predictions" (test set), representing less than 8% of the whole range of values (140.0), which would allow a reasonable degree of predictivity. A plot comparing the predicted and experimental values of the test set is presented in Fig. S1 of the ESI.†

Table 3 Subgroup of eight solvents extracted from the total amount of solvents in order to create the new 54 solvents training set

Solvent	E_T^N	η (cP)	b.p. (°C)
200	0.690	35.140	221
104	0.480	—	208
3i03F	0.590	6.900	176
5F05F	0.699	—	204
113i	—	1.030	170
414t	0.141	—	234
4t13F	0.373	2.140	185
3F23F	0.595	—	171

Table 4 Linear regression parameters from eqn (2)–(4) obtained with the training set of solvents[a]

	E_T^N	η	b.p.
$b_0 \pm e_0$	0.196 ± 0.039	—[b]	111.4 ± 17.6
$b_1 \pm e_1$	0.071 ± 0.012	14.19 ± 4.59	11.7 ± 2.0
$b_2 \pm e_2$	0.200 ± 0.024	—	25.9 ± 5.4
$b_3 \pm e_3$	−0.018 ± 0.005	—	−3.0 ± 1.4
N	39	13	54
R^2	0.953	0.791	0.782
$\sigma(y)$	0.0456	8.16	11.9
F[c]	238.24	45.31	59.64

[a] b_i are the coefficients for each regressions, e_i is the tolerance for the b_i value in a 95% confidence interval. N is the number of cases (solvents data) used in each regression, R^2 is the determination coefficient. [b] As the b_0 coefficient turned out to be non-significant in the standard MLR analysis, fitting was done by forcing the equation to pass through the origin of coordinates. A slight improvement in R^2 was obtained in this way. [c] $F_{(3,\ 35,\ 0.05)} = 2.88$; $F_{(1,\ 12,\ 0.05)} = 4.75$; $F_{(3,\ 50,\ 0.05)} = 2.79$. All equations are statistically significant ($p > 95\%$).

Partial least squares (PLS) regression with topological indices. One problem when using topological indices is the high pair-correlation existing between many of them, given that they often recover similar structural features of the target molecule. This can have undesirable consequences in MLR analyses, since the real significance of a variable cannot be ascertained if it is highly correlated with another one. For instance, when examining variable coefficients in eqn (2) one should be aware that HBA and SC_3^p have a pair correlation coefficient as high as 0.828 (full pair-wise correlation data are gathered in Table S3 in the ESI†).

A possible solution to this problem is to transform the original variables into a new set of a few new orthogonal (not correlated) variables, gathering most of the total variance of data. In the case of PLS regression,[38,39] both the dependent (y) and the independent (x) variables are projected in a new space, trying to maximize the explanation of the variance of y through the variance of latent variables x. Once this relationship is found, the PLS coefficients are projected back to the original x-space, to obtain the corresponding regression coefficients.

When the PLS regression technique was applied to our problem, slightly better models were obtained for two of the three properties considered. The corresponding coefficients and PLS parameters are shown in Table 5, the most important

Table 5 PLS regression results obtained in the treatment of the experimental solvent properties studied in this work

b_i	$E_T^{N\,a}$	η^b	b.p.^{c}
HBA	0.0141	0.238	8.9
HBD	**0.1370**	**9.387**	**46.7**
RB	0.0097	0.713	6.4
φ	−0.0104	−0.001	−2.4
Bal_{JX}	**−0.1500**	**−3.210**	**31.1**
Bal_{JY}	**−0.0806**	**−2.638**	**35.4**
W_r	0.0000	0.003	0.0
Z	0.0007	0.005	−0.2
κ_1^α	0.0026	−0.024	2.1
κ_2^α	−0.0093	0.004	−2.0
κ_3^α	0.0228	**−1.118**	−5.4
SC_0^p	0.0030	−0.013	2.3
SC_1^p	0.0030	−0.013	2.3
SC_2^p	0.0022	0.025	−1.7
SC_3^p	−0.0006	0.136	−6.4
SC_3^{cl}	0.0040	0.105	−7.7
χ_0	0.0041	0.000	1.1
χ_1	0.0044	−0.034	**10.8**
χ_2	0.0100	−0.106	−5.0
χ_3^p	−0.0011	0.819	0.9
χ_3^{cl}	0.0211	−0.188	2.6
$\chi_0^{v.m.}$	−0.0185	−0.560	**−11.2**
$\chi_1^{v.m.}$	−0.0201	−0.344	**−12.8**
$\chi_2^{v.m.}$	−0.0242	**−1.185**	**17.0**
$\chi_3^{p,v.m.}$	**−0.0796**	0.445	**109.8**
$\chi_3^{cl,v.m.}$	**−0.0632**	**−3.142**	**12.0**
b_0	0.9997	30.973	−111.3
N	46	17	62
R^2	0.969 (0.954)^d	0.700 (0.535)	0.891 (0.770)
$\sigma(y)$	0.036	7.29	8.1

[a] PLS regression used 4 latent variables built from the 26 original ones. [b] PLS regression used 3 latent variables built from the 26 original ones. [c] PLS regression used 7 latent variables built from the 26 original ones. [d] Values in parentheses correspond to full cross-validated analyses, *i.e.* each value is predicted by the equation obtained leaving out that solvent. The resulting fitting is therefore more representative of the true predictive ability of the model.

coefficients corresponding again to the hydrogen-bonding indices. Plots of predicted *vs.* experimental values of the properties are displayed in Fig. 3. As can be seen in these plots, in the case of E_T^N the PLS model fits very well the values of most of the 62 solvents used in the analysis. The MUE of the fitted values is 0.028, identical to that obtained in the previous MLR analysis. The full cross-validated predictions (*i.e.*, those performed by leaving the predicted point out of the PLS calculation of the coefficients) are close to normal predictions in all but one case (7F07F), which points to the robustness of the model and the reliability of the predictions. The MUE in this case is only slightly higher, 0.034. On the other hand, viscosity displays a bad behaviour concerning the PLS analysis, with a determination coefficient (R^2) even lower than that found in the MLR analysis. Again, hydrogen bond donor ability and κ_3^α are the topological variables with higher coefficients. However the MUE of the fitted values is 5.86 and that of the cross-validated values increases to 7.88, values which are not far from those obtained in the MLR analyses, although they seem to be too high to allow reliable quantitative predictions.

Finally, the fitting of boiling points is slightly better with the PLS approach (higher R^2 and lower $\sigma(y)$), and the resulting

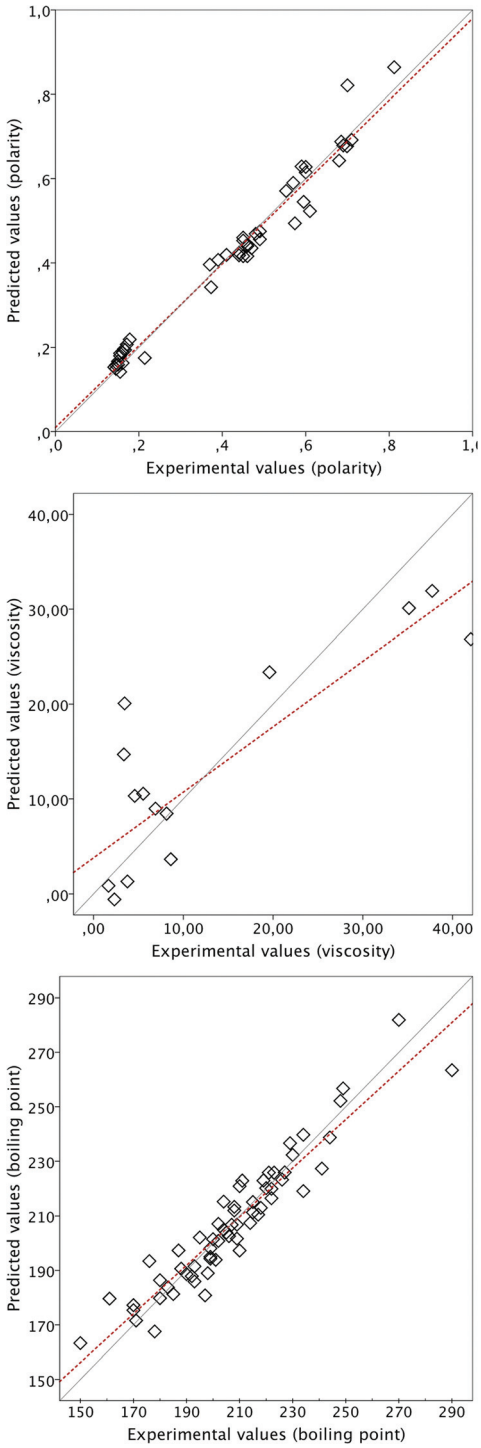


Fig. 3 Plots of predicted *vs.* experimental values of E_T^N (a), viscosity (b), and boiling point (c), as calculated through PLS analysis using topological indices.

model is quite robust, with only three outliers: glycerol itself (000), 444 and 7F07F. In this case, the MUE are 6.2 (fitted values) and 8.4 (cross-validated values), slightly better than those found in the MLR analyses.

In order to have more reliable proof of the predictive ability of these equations, we split the data again into the same training and tests sets used in the MLR analyses. The results of the

Table 6 PLS regression results obtained in the treatment of the training set of solvents

b_i	$E_T^{N\,a}$	η^b	b.p.^{c}
HBA	0.0145	0.142	4.181
HBD	**0.1390**	**10.982**	**38.876**
RB	0.0097	0.717	10.645
φ	−0.0096	0.047	1.938
Bal_{JX}	**−0.1680**	**−2.110**	**31.450**
Bal_{JY}	**−0.0990**	**−2.036**	**29.459**
W_r	0.0000	0.002	−0.006
Z	0.0006	0.003	−0.226
κ_1^α	0.0022	−0.019	−0.010
κ_2^α	−0.0088	0.047	2.044
κ_3^α	0.0213	**−1.290**	−1.801
SC_0^p	0.0027	−0.011	0.155
SC_1^p	0.0027	−0.011	0.155
SC_2^p	0.0022	0.019	−1.142
SC_3^p	−0.0006	0.125	−3.202
SC_3^{cl}	0.0042	0.079	−1.859
χ_0	0.0038	−0.001	−0.943
χ_1	0.0034	−0.017	3.560
χ_2	0.0101	−0.128	−4.585
χ_3^p	−0.0008	0.757	−1.390
χ_3^{cl}	0.0219	−0.268	7.733
$\chi_0^{v.m.}$	−0.0180	−0.441	−8.287
$\chi_1^{v.m.}$	−0.0184	−0.186	−7.186
$\chi_2^{v.m.}$	−0.0198	−0.864	−0.827
$\chi_3^{p,v.m.}$	**−0.0699**	0.959	**69.209**
$\chi_3^{cl,v.m.}$	**−0.0519**	**−3.170**	**19.237**
b_0	1.0874	22.028	−56.840
N	39	13	54
R^2	0.967	0.668	0.874
$\sigma(y)$	0.036	7.55	8.7

[a] PLS regression used 4 latent variables built from the 26 original ones. [b] PLS regression used 3 latent variables built from the 26 original ones. [c] PLS regression used 8 latent variables built from the 26 original ones.

corresponding regressions are shown in Table 6. Plots of experimental *vs.* predicted values (including solvents in the test set) are shown in Fig. S2 (ESI†).

As can be seen from the values in Table 6, both the goodness of the fitting and the regression coefficients obtained with the training set of solvents are quite similar to those calculated with the full set.

Concerning the prediction errors, the MUE values for E_T^N are 0.027 for the training set (almost identical to that calculated with the full set of solvents) and 0.030 for the test set, which points to a good predictivity of the equations developed. Concerning the viscosity, the corresponding MUE values are 6.33 and 4.62 for the training and test sets, respectively, which are also quite close to that obtained with the full set of solvents (5.86) and point to a worse predictivity of this property by the model developed. Finally, the MUE values for the prediction of boiling points are 6.6 (training set) and 10.3 (test set). Even if the latter is clearly higher, it still represents about 7% of the full range of b.p. values, which may be enough to obtain a reasonable predictivity of this solvent property.

Multiple linear regression (MLR) with DARC/PELCO descriptors. In this case we used again the stepwise method to include in the regression equation only those variables which are statistically significant. It should be noted that for predictive purposes, given the local character of the DARC/PELCO descriptors, the values of the coefficients of all the variables not included in the final equations must be taken as zero. The three MLR equations thus obtained are the following:

$$E_T^N = b_0 + b_1 \cdot A_1 + b_2 \cdot B_{F2} + b_3 \cdot A_2 + b_4 \cdot B_2 + b_5 \cdot C_{F2} + b_6 \cdot C_2 \quad (5)$$

$$\eta = b_0 + b_1 \cdot A_2 + b_2 \cdot C_{F2} + b_3 \cdot A_1 \quad (6)$$

$$\mathrm{bp} = b_0 + b_1 \cdot D_2 + b_2 \cdot A_2 + b_3 \cdot C_1 + b_4 \cdot C_2 + b_5 \cdot B_2 + b_6 \cdot C_{F2} + b_7 \cdot B_{F2} + b_8 \cdot A_1 \quad (7)$$

Table 7 Linear regression parameters from eqn (5)–(7)

b_i	E_T^N	η	b.p.
$b_0 \pm e_0$	0.851 ± 0.057	70.79 ± 5.45	278.2 ± 10.6
$b_1 \pm e_1$	−0.278 ± 0.023	−32.50 ± 3.10	19.1 ± 3.60
$b_2 \pm e_2$	0.140 ± 0.024	6.90 ± 2.40	−55.6 ± 6.68
$b_3 \pm e_3$	−0.160 ± 0.038	−3.52 ± 2.50	33.6 ± 6.32
$b_4 \pm e_4$	−0.026 ± 0.012	—	12.6 ± 2.49
$b_5 \pm e_5$	−0.059 ± 0.032	—	7.9 ± 1.86
$b_6 \pm e_6$	−0.016 ± 0.014	—	12.0 ± 5.84
$b_7 \pm e_7$	—	—	7.0 ± 3.98
$b_8 \pm e_8$	—	—	−6.1 ± 3.97
N	46	17	62
R^2	0.972	0.981	0.933
$\sigma(y)$	0.036	2.08	6,9
F^a	229.23	228.29	92.18

[a] $F_{(6,\ 39,\ 0.05)} = 2.34$; $F_{(3,\ 13,\ 0.05)} = 3.34$; $F_{(8,\ 53,\ 0.05)} = 2.18$. All equations are statistically significant ($p > 95\%$).

The corresponding coefficient values and MLR parameters are shown in Table 7, and the plots of predicted *vs.* experimental values of the properties are displayed in Fig. 4.

As can be seen, the fitting of the three properties is better than those described with the precedent approaches. Even the viscosity displays good values. In a first approach, this cannot be ascribed to overfitting, given that the final equation has only three independent variables to fit seventeen data, *i.e.* more than five times data than variables. Similarly, boiling point also displays a very good fitting, with a low standard error (*ca.* 7 °C).

The robustness of the method was tested again by removing the same test set of solvents (Table 3) from the entire data and, as can be seen from the values gathered in Table 8, the regression coefficients in eqn (5)–(7) do not change dramatically, all values lying within the calculated confidence margins.

Fig. S3 (in ESI†) shows the predicted data for the eight members of the test group. It can be seen that the best predictable property is the boiling point, whose deviations from experiment are less than ten percent in the worst case. E_T^N displays a more erratic behaviour, especially in the case of fluorinated compounds, for which deviations are important in relative terms, although they preserve the qualitative order experimentally observed. As expected, the largest deviations correspond to those structural features less represented in the

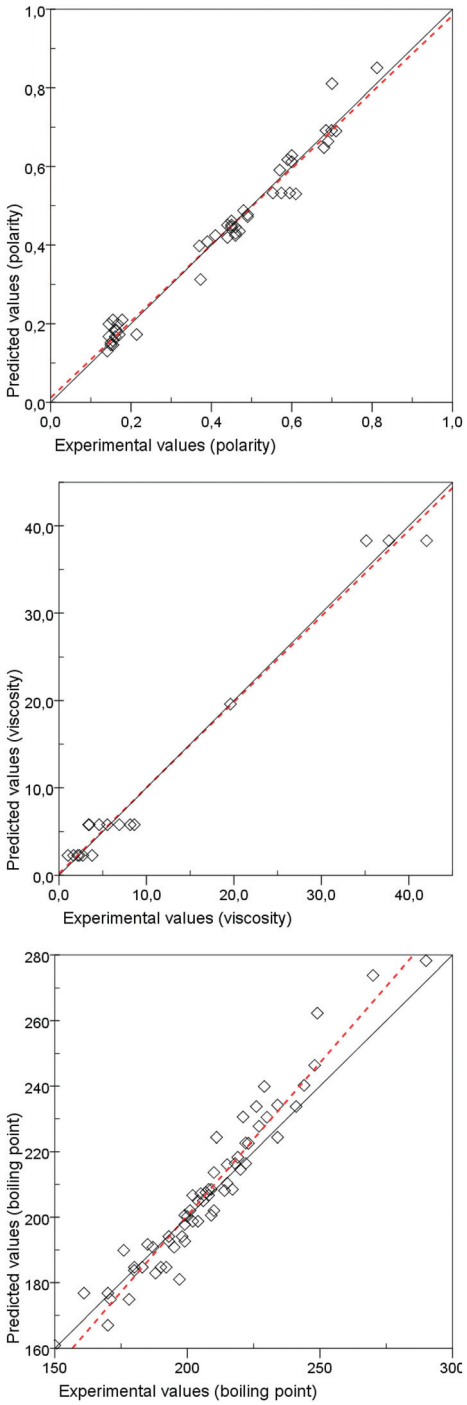


Fig. 4 Plots of predicted *vs.* experimental values of E_T^N (a), viscosity (b), and boiling point (c), as calculated through MLR analysis using DARC/PELCO descriptors.

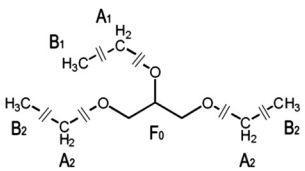
training set (highly branched and highly fluorinated chains). Concerning the MUE, in all cases the values for the fitted values using the solvent training set are lower than those obtained using the topological descriptors (0.024, 1.37 and 4.6 for E_T^N, viscosity and b.p., respectively), but these values are significantly higher for the test set (0.051, 2.05 and 10.3, respectively). Anyway, these errors represent between 5% and 8% of the full range of values, which point to a reasonably good predictivity of these equations.

Table 8 Linear regression factors from eqn (5)–(7) using a reduced training set of 54 solvents

b_i	E_T^N	η	b.p.
$b_0 \pm e_0$	0.839 ± 0.059	74.13 ± 6.81	281.3 ± 11.0
$b_1 \pm e_1$	−0.289 ± 0.025	−34.26 ± 3.78	18.4 ± 3.5
$b_2 \pm e_2$	0.126 ± 0.026	6.99 ± 2.50	−56.4 ± 6.8
$b_3 \pm e_3$	−0.148 ± 0.039	−2.99 ± 2.99	33.0 ± 6.1
$b_4 \pm e_4$	−0.030 ± 0.013	—	12.2 ± 2.4
$b_5 \pm e_5$	−0.055 ± 0.041	—	7.8 ± 1.9
$b_6 \pm e_6$	−0.016 ± 0.014	—	10.6 ± 7.5
$b_7 \pm e_7$	—	—	8.1 ± 4.3
$b_8 \pm e_8$	—	—	−6.5 ± 4.1
N	39	13	54
R^2	0.975	0.983	0.940
$\sigma(y)$	0.035	2.05	6.6
F^a	206.40	174.86	87.83

[a] $F_{(6,\ 32,\ 0.05)} = 2.42$; $F_{(3,\ 9,\ 0.05)} = 3.71$; $F_{(8,\ 45,\ 0.05)} = 2.18$. All equations are statistically significant ($p > 95\%$).

Table 9 Example of boiling point prediction of 1,2,3-triethoxypropane (222) using the linear regression obtained with DARC/PELCO descriptors[a]



	b_i	No. of fragments	Total contribution
F_0	278.2	1	278.2
A_1	−6.1	1	−6.1
A_2	−55.6	2	−111.2
B_1	0.0	1	0.0
B_2	7.9	2	15.8
			176.7

[a] Experimental value: 181 °C.[40]

As already mentioned, DARC/PELCO descriptors are highly intuitive, given their straightforward matching with the molecular structure. As a consequence, the prediction of the property of a new compound is extremely simple. As an example we present the calculation of the boiling point of a glycerol-derived solvent, not belonging to our 62 solvent set, namely 1,2,3-triethoxypropane (222). This compound and its boiling point were described in the literature, so the example represents a "real world" prediction, given that the property was determined by other authors using a different experimental technique. Table 9 gathers the detailed prediction procedure from the calculated regression coefficients. As can be seen, the predicted value (177 °C) is reasonably close to the experimental one (181 °C),[40] and within the standard regression error (*ca.* 95% predicted values should be within a range of ±14 °C from experimental ones).

Table 10 Linear regression factors from eqn (8)–(10)

b_i	E_T^N	η	b.p.
$b_0 \pm e_0$	0.523 ± 0.122	67.55 ± 3094	292.6 ± 35.8
$b_1 \pm e_1$	0.140 ± 0.042	−35.86 ± 2.51	−49.7 ± 9.0
$b_2 \pm e_2$	0.177 ± 0.020	−5.27 ± 1.75	12.9 ± 3.4
$b_3 \pm e_3$	−0.099 ± 0.043	0.99 ± 0.23	−26.0 ± 13.2
$b_4 \pm e_4$	−0.026 ± 0.010		20.4 ± 6.9
$b_5 \pm e_5$			−8.3 ± 7.5
N	46	17	62
R^2	0.968	0.989	0.932
$\sigma(y)$	0.036	1.46	6.8
F^a	314.31	376.64	153.47

[a] $F_{(4,\ 41,\ 0.05)} = 2.34$; $F_{(3,\ 13,\ 0.05)} = 3.34$; $F_{(5,\ 56,\ 0.05)} = 2.18$. All equations are statistically significant ($p > 95\%$).

Multiple linear regression (MLR) with mixed DARC/PELCO and topological descriptors. More compact prediction equations (eqn (8)–(10)) were obtained by mixing DARC/PELCO and topological indices, thus considering simultaneously local and global structure descriptors, respectively. The coefficients and statistical parameters for these regressions are gathered in Table 10, and the plots of predicted *vs.* experimental values of the properties are displayed in Fig. 5.

$$E_T^N = b_0 + b_1 \cdot \mathrm{HBD} + b_2 \cdot B_{F2} + b_3 \cdot A_1 + b_4 \cdot \chi_0^{v.m.} \quad (8)$$

$$\eta = b_0 + b_1 \cdot A_2 + b_2 \cdot A_1 + b_3 \cdot \chi_0 \quad (9)$$

$$\mathrm{bp} = b_0 + b_1 \cdot A_2 + b_2 \cdot \mathrm{RB} + b_3 \cdot \mathrm{Bal}_{JY} + b_4 \cdot \chi_2^{v.m.} + b_5 \cdot \chi_0^{v.m.} \quad (10)$$

Although the statistical tests are very similar to those obtained with the DARC/PELCO descriptors only, less independent variables are used in the final equations, leading to higher ratios of the number of cases to the number of variables. In the case of viscosity, the number of independent variables does not change, but the standard error of the predictions is slightly improved (from 2.05 to 1.46 cP).

The robustness and predictivity of these equations were again tested by splitting the solvent set into training and test sets. The corresponding regression results are presented in Table 11. As can be seen, there are no significant changes in fitting parameters and regression coefficients. Fig. S4 (in ESI†) shows the predicted data for the eight members of the test group.

The comparison of the MUE calculated with the fitting of the training set and the test set indicates that prediction errors are significantly higher in the latter case, but they are anyway lower than those obtained with the precedent models, representing 5–6% of the full range of experimental values in all cases. A summary of the MUE calculated for all the equations developed in this work is presented in Table 12.

If we take the MUE calculated for the test set as a measure of the actual predictivity of the equations, we can conclude that good predictive models have been developed for the three properties under study. Topological descriptors seem to be more adequate for the prediction of E_N^T, mostly due to the poor predictions of DARC/PELCO descriptors for fluorinated solvents. The latter, on the other hand, perform much better in the prediction of viscosities. Overall, the mixed DARC/PELCO-topological model constitutes the best compromise for reasonably predicting the three solvent properties studied here.

A referee suggested that PLS analyses could also be applied to the DARC/PELCO and mixed parameter models. The

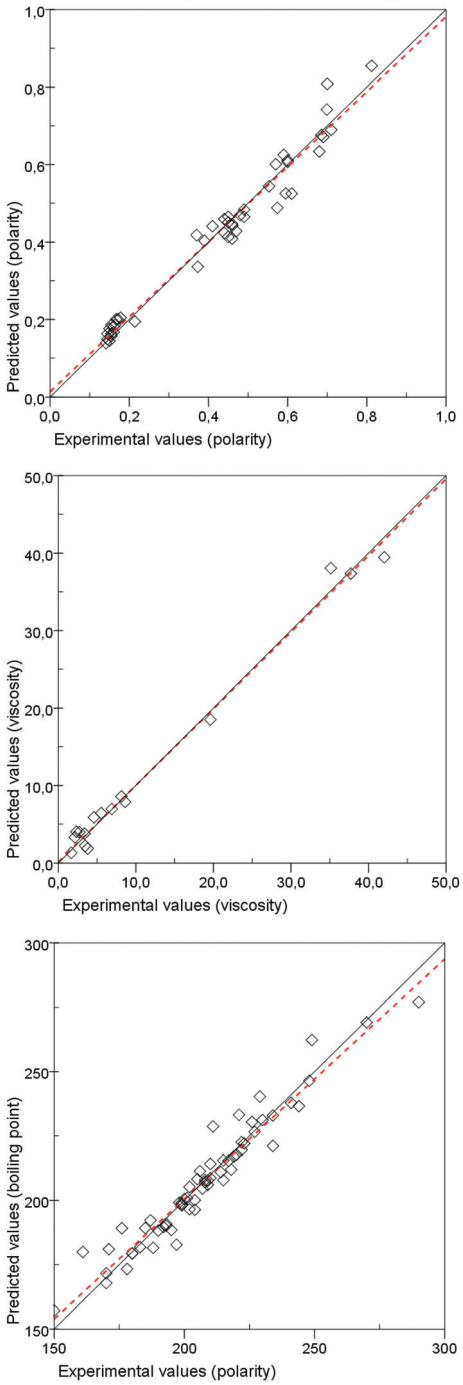


Fig. 5 Plots of predicted *vs.* experimental values of E_T^N (a), viscosity (b), and boiling point (c), as calculated through MLR analysis using topological indices and DARC/PELCO descriptors.

Table 11 Linear regression factors from eqn (8)–(10)

b_i	E_T^N	η	bp
$b_0 \pm e_0$	0.512 ± 0.129	70.41 ± 4.49	288.6 ± 38.2
$b_1 \pm e_1$	0.139 ± 0.045	−37.83 ± 2.71	−49.3 ± 8.9
$b_2 \pm e_2$	0.173 ± 0.021	−5.46 ± 1.91	13.3 ± 3.5
$b_3 \pm e_3$	−0.112 ± 0.050	1.03 ± 0.23	−22.5 ± 14.1
$b_4 \pm e_4$	−0.024 ± 0.011		19.5 ± 6.6
$b_5 \pm e_5$			−9.3 ± 7.6
N	39	13	54
R^2	0.969	0.993	0.944
$\sigma(y)$	0.038	1.35	6.2
F^a	268.19	405.64	161.36

[a] $F_{(4,\ 34,\ 0.05)} = 2.64$; $F_{(3,\ 9,\ 0.05)} = 3.71$; $F_{(5,\ 48,\ 0.05)} = 2.42$. All equations are statistically significant ($p > 95\%$).

Table 12 Mean unsigned errors (MUE) calculated for the different equations developed in this work[a]

	E_T^N		η		b.p.	
Model	Training	Test	Training	Test	Training	Test
MLR topol.	0.028	**0.030**	7.28	5.08	8.6	*10.9*
PLS topol.	0.027	**0.030**	6.34	4.62	6.6	*10.3*
MLR D.-P.	0.024	*0.051*	1.37	**2.05**	4.6	*10.3*
MLR mixed	0.026	**0.033**	1.02	**1.95**	3.9	*8.8*

[a] Boldface values indicate errors within 5% of the full range of experimental values, and italicized values indicate errors within 8% of the full range.

corresponding results can be found in the ESI,† but in no case improvement over the MLR equations could be obtained, so they will not be discussed here.

Experimental

Glycerol-based solvents were obtained by ring opening of either the appropriate glycidol ether (non-symmetric glycerol-based solvents) or epichlorohydrin (symmetric glycerol-based solvents) with the corresponding alkoxide in alcoholic media, and purified by vacuum distillation as described previously.[1]

The complete list of the 62 solvents used in QSPR analyses and the values of the experimental properties studied is presented in Table 1.

Different topological descriptors were calculated for the molecular structures of every solvent using *Materials Studio Modeling 4.0* from Accelrys. This software can calculate topological descriptors on the basis of molecular structural information. All these descriptors are gathered in Table S1 of the ESI.†

DARC/PELCO descriptors were generated from the scheme shown in Fig. 1. The presence of a C unit (bearing the corresponding hydrogen atoms) was codified as 1 in the data matrix (2 if the unit is simultaneously present at both symmetric sides of the glycerol moiety). C units bearing fluorine atoms were codified as independent variables (those starting with "F" in the regression analyses). The final DARC/PELCO matrix is gathered in Table S2.†

Multiple linear regression analyses were carried out using SPSS software. In all the tables the following information is provided:

– Regression coefficients b_i, as defined previously ($\boldsymbol{B} = (\boldsymbol{X}^T \cdot \boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$).

– Individual confidence intervals (at the 95% probability level) of each b_i coefficient. These confidence intervals are calculated from the estimated standard error of b_i and the Student's test with $N - p$ degrees of freedom:

$$e_i = \text{s.e.}(b_i) \cdot t(N - p, 0.975)$$

– The number of cases included in the regression, N.

– Multiple determination coefficient, R^2, which is a measure of the proportion of the total variation about the mean of y explained by the regression.

– Standard error of the regression $\sigma(y)$ is the root square of the residual mean square, and it is the estimate of the error with which any observed value of y could be predicted by the regression equation.

– F value, defined as the quotient of the regression and residual mean squares. When compared with a Fisher–Snedecor F distribution with $p - 1$ and $N - p$ degrees of freedom, at a 95% probability level (values given in the footnotes of the tables), it allows establishing whether the variance explained by the regression equation is significantly different from that of the error. More strictly, it tests the H_0 hypothesis, *i.e.*, that all regression coefficients are zero. If the calculated F value is larger than the tabulated one, the hypothesis is rejected, and the equation is considered statistically significant.

– The stepwise linear regression procedure is a method to select the "best" regression equation from a set of independent variables, x. Each variable is sequentially included in the equation, following its single correlation with the response, y. For each new variable entering, a partial F-test is performed to see whether the improvement in the equation is significant. If the variable is accepted, then partial F-tests are also performed for the rest of variables already in the equation. Those not passing the test are then eliminated. The procedure is repeated until no more variables are included in the equation. Partial F-tests are carried out at a 90% probability level.

The mean unsigned (or absolute) error (MUE or MAE) is an average of the absolute errors $e_i = |\hat{y}_i - y_i|$, where $\hat{y}_i$ is the value predicted by the model and y_i the experimental value.

Conclusions

In this study three characteristic properties relevant to classify solvents and facilitate the search of substitution uses have been investigated in a series of 62 glycerol derivatives that can be used as solvents. Global topological descriptors, based on the molecular graphs, have been successfully applied to analyze and predict solvent polarities, both using traditional

MLR and PLS regression analyses. However, boiling points and viscosities are not so well modeled using this kind of structural variables.

On the other hand, DARC/PELCO local structural descriptors have been revealed to be clearly superior to describe the viscosity of this family of solvents. Boiling points are similarly well predicted with both kinds of approaches.

Overall, the mixed model with DARC/PELCO and topological descriptors constitutes the best compromise for reasonably predicting the three solvent properties studied in this work.

Highly significant regression equations have been developed for the three properties under study. The robustness and predictive value of these equations have been demonstrated through the use of an independent test set of solvents. Therefore, the QSPR models developed provide significant additional insight into the relationship between the molecular structure and some fundamental solvent properties.

Based on these results, it seems that quantitative structure–activity/property relationships (QSAR/QSPR) could be quite useful for *in silico* prediction of physico-chemical properties, allowing a faster selection of target solvents for a given application.

Acknowledgements

Financial support from the Spanish MINECO (project CTQ2011-28124-C02-01), the European Social Fund (ESF) and the Gobierno de Aragón (Grupo Consolidado E11) is gratefully acknowledged.

Notes and references

1 J. I. García, H. García-Marín, J. A. Mayoral and P. Pérez, *Green Chem.*, 2010, **12**, 426.
2 R. D. Cramer, *J. Am. Chem. Soc.*, 1980, **102**, 1837.
3 R. Carlson, *Design and Optimization in Organic Synthesis*, Elsevier, Amsterdam, 1992.
4 M. Chastrette, M. Rajzmann, M. Chanon and K. F. Purcell, *J. Am. Chem. Soc.*, 1985, **107**, 1.
5 I. A. Koppel and V. A. Palm, The influence of the solvent on organic reactivity, in *Advances in Linear Free Energy Relationships*, Plenum Press, London, 1972.
6 M. J. Kamlet, J. L. Abboud, M. H. Abraham and R. W. Taft, *J. Org. Chem.*, 1983, **48**(17), 2877.
7 J. Catalán, *Solvent Effects based on non-HBD Solvents in Handbook of Solvents*, William Andrew Publishing, New York, 2001.
8 C. Reichardt, *Solvents and Solvent Effects*, Wiley-VCH, Weinhein, 3th edn, 2003.
9 M. Ravi, A. J. Hopfinger, R. E. Hormann and L. Dinan, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1587.
10 B. T. Luke, *J. Mol. Struct. (THEOCHEM)*, 1999, **13**, 468.
11 P. Bruneau, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1605.
12 A. R. Katritzky, R. Petrukhin and D. Tatham, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 679.
13 J. Ghasemi, S. Saaidpour and S. D. Brown, *J. Mol. Struct. (THEOCHEM)*, 2007, **805**, 27.
14 N. Brauner, M. Shachamb, G. S. Cholakovc and R. P. Statevad, *Chem. Eng. Sci.*, 2005, **60**, 5458.
15 P. Lind, C. Lopes, K. Oberg and B. Eliasson, *Chem. Phys. Lett.*, 2004, **387**, 238.
16 P. Ungerer, C. Nieto-Draghi, B. Rousseau, G. Ahunbay and V. Lachet, *J. Mol. Liq.*, 2007, **134**, 71.
17 J. B. Ghasemi, A. Abdolmaleki and N. Mandoumi, *J. Hazard. Mater.*, 2009, **161**, 74.
18 M. H. Fatemi and M. Haghdadi, *J. Mol. Struct.*, 2008, **886**, 43.
19 J. S. Torrecilla, J. Palomar, J. Lemus and F. Rodríguez, *Green Chem.*, 2010, **12**, 123.
20 M. Alvarez-Guerra and A. Irabien, *Green Chem.*, 2011, **13**, 1507.
21 F. Yan, S. Xia, Q. Wang and P. Ma, *J. Chem. Eng. Data*, 2012, **57**, 2252.
22 V. Consonni, R. Todeschini, M. Pavan and P. Gramatica, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 693.
23 G. Krenkel, E. A. Castro and A. A. Toropov, *J. Mol. Struct. (THEOCHEM)*, 2001, **542**, 107.
24 J. Ghasemi, S. Shahmirani and E. V. Farahani, *Ann. Chim.*, 2006, **96**, 327.
25 L. B. Kier and L. H. Hall, *Molecular Connectivity in Structure–Activity Analysis*, Research Studies Press Ltd, New York, 1985.
26 A. R. Katritzky, D. C. Fara, M. Kuanar, E. Hur and M. Karelson, *J. Phys. Chem. A*, 2005, **109**, 10323.
27 N. R. Draper and H. Smith, *Applied Regression Analysis*, Wiley-Interscience, 1998.
28 K. Dimroth, C. Reichardt, T. Siepmann and F. Bohlmann, *Liebigs Ann. Chem.*, 1963, **1**, 661.
29 K. Dimroth, C. Reichardt and A. Schweig, *Liebigs Ann. Chem.*, 1963, **95**, 669.
30 D. R. Lide, *Handbook of Chemistry and Physics*, CRC, New York, 84th edn, 2004.
31 A. R. Katritzky and E. V. Gordeeva, *J. Chem. Inf. Comput. Sci.*, 1993, 835.
32 L. H. Hall and L. B. Kier, *Rev. Comput. Chem. II*, 1991, 367.
33 A. T. Balaban, *Chem. Phys. Lett.*, 1982, 309.
34 H. Wiener, *J. Chem. Phys.*, 1947, 17.
35 D. Bonchev, *Information Theoretic Indices for Characterization of Chemical Structures*, Research Studies Press Ltd., New York, 1983.
36 L. B. Kier and L. H. Hall, *Molecular Connectivity Indices in Chemistry and Drug Research*, deStevens, New York, 1976.
37 J. E. Dubois, *Computer Representation and Manipulation of Chemical Information*, Wiley, New York, 1974.
38 S. Wold, A. Ruhe, H. Wold and W. Dunn, *SIAM J. Sci. Stat. Comput.*, 1984, **5**, 735.
39 P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 1986, **185**, 1.
40 A. Fairbourne, G. P. Gibson and D. W. Stephens, *J. Chem. Soc.*, 1931, 445.