Development of Quantitative Structure−Property Relationships Using Calculated Descriptors for the Prediction of the Physicochemical Properties (n_D, ρ, bp, ϵ, η) of a Series of Organic Solvents

Marina Cocchi,* Pier Giuseppe De Benedetti, Renato Seeber, Lorenzo Tassi, and Alessandro Ulrici

Department of Chemistry, University of Modena, via G. Campi 183, 41100 Modena, Italy

Received June 16, 1999

Quantitative structure−property relationship (QSPR) models were derived for predicting boiling point (at 760 mmHg), density (at 25 °C), viscosity (at 25 °C), static dielectric constant (at 25 °C), and refractive index (at 20 °C) of a series of pure organic solvents of structural formula $X-CH_2CH_2-Y$. A very large number of calculated molecular descriptors were derived by quantum chemical methods, molecular topology, and molecular geometry by using the CODESSA software package. A comparative analysis of the multiple linear regression techniques (heuristic and best multilinear regression) implemented in CODESSA, with the multivariate PLS/GOLPE method, has been carried out. The performance of the different regression models has been evaluated by the standard deviation of prediction errors, calculated for the compounds of both the training set (internal validation) and the test set (external validation). Satisfactory QSPR models, from both predictive and interpretative point of views, have been obtained for all the studied properties.

INTRODUCTION

The correlation and prediction of physicochemical properties of pure liquids and of mixtures, such as boiling point, density, viscosity, static dielectric constant, and refractive index, is of practical (process design and control) and theoretical (role of the molecular structure in determining the macroscopic properties of the solvent) relevance to both chemists and engineers. Traditionally, procedures for estimating these properties have been based either on theoretical relationships often making use of empirical parameters that have to be fitted or on empirical relationships derived from additive−constitutive schemes based on atomic groups or bonds contribution within the molecule.[1−5] More recently, the quantitative structure−property relationships (QSPR) approach[6,7] based on calculated molecular descriptors has been applied especially to predict boiling points,[8,9] partition coefficients,[10,11] chromatographic retention indexes,[12−14] surface tension, and critical temperatures,[15,16] while only a few studies have dealt with viscosity,[1,17,18] refractive index,[19] and static dielectric constant.[1,20] The use of calculated molecular descriptors in QSPR analysis has two main advantages: (a) the descriptors can be univocally defined for any molecular structure or fragment; (b) thanks to the high and well-defined physical information content encoded in many theoretical descriptors (for example, those derived from quantum chemical theory), they can clarify the mechanism relating the studied property with the chemical structure. Moreover, QSPR models based on calculated descriptors help understanding of the inter- and intramolecular interactions that are mainly responsible for the behavior of complex chemical systems and processes.

In the present study, we derive QSPR models capable of giving account for the following properties: boiling point (at 760 mmHg), density (at 25 °C), viscosity (at 25 °C), static dielectric constant (at 25 °C), and refractive index (at 20 °C) of a series of pure organic solvents. This constitutes a preliminary step toward the modeling of the thermophysical properties of mixtures, which have been experimentally fully characterized by our research group.[21] The compounds considered have structural formula $X-CH_2CH_2-Y$, where the X and Y fragments, including alkyls, aromatics, halogens, electron acceptors and donors, and hydrogen-bonding groups, were chosen to span over different stereoelectronic features and constitute a structural heterogeneous data set. The property data were collected from the literature; depending on the available data for each property, the number of compounds in the training set varies from 23 to 67 and that in the test set (used for external validation of the QSPR models) varies from 9 to 29. Despite the fact that the number of compounds is limited compared to similar comprehensive studies, the variation range of the various properties is quite acceptable.

A very large number of calculated molecular descriptors was derived by quantum chemical methods, molecular topology, and molecular geometry, by using the CODESSA software package.[6,22]

A significant aspect in QSPR is whether and how to select, from a very large number of calculated indexes, a limited set of descriptors that represents the best choice for predicting and accounting for the property data; several approaches to this task have recently been reported in the literature.[23−27] Despite the fact that multivariate data analysis techniques can handle many collinear variables, there are two main reasons for operating a variables selection procedure when using theoretical indexes: (a) collinearity among descriptors indicates that their physicochemical meaning and information content may be the same, i.e., collinear descriptors are interchangeable; (b) the interpretation of the derived QSPR

* Corresponding author: (fax) (+39) 059/373543; (e-mail) cocchi@unimo.it.

10.1021/ci9903298 CCC: $18.00
Published on Web 10/09/1999

Table 1. Properties Data Values for the List of Training Set Compounds: Their Refractive Index, Density, Boiling Point, and Permittivity Data and References (Numbers in Parentheses)

no.	compound	formula	n_D[a]	ρ[b]	bp[c]	ϵ[d]
1	1,4-dioxane	$\mathbf{OCH_2CH_2O}CH_2CH_2$	1.42241 (49)	1.028112 (57)	101.320 (49)	3.208 (48)
2	ethanol	$\mathbf{HO}CH_2CH_2\mathbf{H}$	1.36139 (49)	0.78506 (49)	78.325 (49)	24.30 (49)
3	ethane-1,2-diol	$\mathbf{HO}CH_2CH_2\mathbf{OH}$	1.4318 (49)	1.109913 (58)	197.85 (49)	40.97 (50)
4	2-methoxyethanol	$\mathbf{HO}CH_2CH_2\mathbf{OCH_3}$	1.4017 (49)	0.960288 (59)	124.4 (49)	16.93 (51)
5	1,2-dimethoxyethane	$\mathbf{CH_3O}CH_2CH_2\mathbf{OCH_3}$	1.3796 (51)	0.861506 (58)	83.5 (54)	7.55 (52)
6	2-chloroethanol	$\mathbf{HO}CH_2CH_2\mathbf{Cl}$	1.44380 (49)[g]	1.2019 (49)[e]	128.6 (49)	25.07 (53)
7	1,2-dichloroethane	$\mathbf{Cl}CH_2CH_2\mathbf{Cl}$	1.44759 (49)[g]	1.245518 (60)	83.483 (49)	10.69 (53)
8	bromoethane	$\mathbf{H}CH_2CH_2\mathbf{Br}$	1.42481 (49)	1.44030 (49)[l]	38.386 (49)	9.39 (49)[e]
9	1,2-dibromoethane	$\mathbf{Br}CH_2CH_2\mathbf{Br}$	1.54160 (49)[g]	2.1701 (49)	131.7 (49)	4.78 (49)
10	iodoethane	$\mathbf{H}CH_2CH_2\mathbf{I}$	1.51369 (49)	1.9358 (49)[e]	72.30 (49)	7.82 (49)[e]
11	phenylethane	$\mathbf{H}CH_2CH_2\mathbf{C_6H_5}$	1.49594 (49)	0.86264 (49)	136.187 (49)	2.412 (49)[e]
12	nitroethane	$\mathbf{H}CH_2CH_2\mathbf{NO_2}$	1.3920 (49)	1.03819 (49)	114.0 (49)	28.06 (49)[l]
13	chloropropane	$\mathbf{CH_3}CH_2CH_2\mathbf{Cl}$	1.38800 (49)	0.87994 (49)[l]	46.60 (49)	7.7 (49)[e]
14	bromopropane	$\mathbf{CH_3}CH_2CH_2\mathbf{Br}$	1.43695 (49)[g]	1.34305 (49)	71.03 (49)	8.09 (49)
15	1-propanol	$\mathbf{CH_3}CH_2CH_2\mathbf{OH}$	1.38556 (49)	0.79950 (49)	97.15 (49)	20.1 (49)
16	1,2-diaminoethane	$\mathbf{NH_2}CH_2CH_2\mathbf{NH_2}$	1.45677 (49)	0.8977 (49)[e]	117.0 (49)	14.2 (49)
17	butan-2-one	$\mathbf{CH_3CO}CH_2CH_2\mathbf{H}$	1.37891 (54)	0.799876 (54)	79.50 (49)	17.64 (54)
18	propanenitrile	$\mathbf{H}CH_2CH_2\mathbf{CN}$	1.36812 (49)[g]	0.77682 (49)	97.20 (49)	27.2 (49)[e]
19	1-iodopropane	$\mathbf{CH_3}CH_2CH_2\mathbf{I}$	1.5041 (49)	1.72997 (49)[l]	102.45 (49)	7 (49)[e]
20	1-nitropropane	$\mathbf{CH_3}CH_2CH_2\mathbf{NO_2}$	1.4016 (49)	0.99546 (49)	131.4 (49)	23.24 (49)[l]
21	butanenitrile	$\mathbf{CH_3}CH_2CH_2\mathbf{CN}$	1.38600 (49)[g]	0.78183 (49)[l]	117.9 (49)	20.3 (49)[i]
22	1-phenylpropane	$\mathbf{CH_3}CH_2CH_2\mathbf{C_6H_5}$	1.49202 (55)	0.85780 (55)	159.217 (55)	2.226 (55)[e]
23	valeronitrile	$\mathbf{CN}CH_2CH_2\mathbf{CH_2CH_3}$	1.3975 (51)	0.7952 (51)	141.56 (51)	20.04 (51)[e]
24	1-fluoro-2-chloroethane	$\mathbf{F}CH_2CH_2\mathbf{Cl}$	1.3775 (54)	1.1683 (51)	52.8 (51)	
25	2-bromoethanol	$\mathbf{Br}CH_2CH_2\mathbf{OH}$	1.4915 (54)	1.7629 (54)[e]	149.5 (54)[m]	
26	1,2-difluoroethane	$\mathbf{F}CH_2CH_2\mathbf{F}$	1.3014 (51)	1.024 (51)	30.7 (54)	
27	2-fluoroethanol	$\mathbf{F}CH_2CH_2\mathbf{OH}$	1.3647 (54)[h]	1.0020 (51)	103.5 (54)	
28	1-methoxypropane	$\mathbf{CH_3}CH_2CH_2\mathbf{OCH_3}$	1.3590 (51)	0.723 (51)[e]	38.5 (54)	
29	2-aminoethanol	$\mathbf{NH_2}CH_2CH_2\mathbf{OH}$	1.4539 (49)	1.01170 (51)	171.1 (49)	
30	3-methoxypropionitrile	$\mathbf{CN}CH_2CH_2\mathbf{OCH_3}$	1.40317 (51)	0.9398 (51)[e]	164 (54)	
31	1-aminopropane	$\mathbf{CH_3}CH_2CH_2\mathbf{NH_2}$	1.38815 (49)	0.7110 (51)	48.5 (49)	
32	2,2-dimethylbutane	$\mathbf{(CH_3)_3}CCH_2CH_2\mathbf{H}$	1.36876 (49)	0.64446 (49)	49.741 (49)	
33	1-chloro-2-methoxyethane	$\mathbf{Cl}CH_2CH_2\mathbf{OCH_3}$	1.4111 (54)	1.0461 (51)	92.5 (54)	
34	1-chloro-2-bromoethane	$\mathbf{Cl}CH_2CH_2\mathbf{Br}$	1.4908 (54)	1.7392 (54)[e]	107 (54)	
35	1-bromo-2-fluoroethane	$\mathbf{Br}CH_2CH_2\mathbf{F}$	1.4236 (54)	1.7044 (54)	71.5 (54)	
36	1-bromo-2-methoxyethane	$\mathbf{Br}CH_2CH_2\mathbf{OCH_3}$	1.44753 (54)	1.4369 (51)	110.3 (54)	
37	2-iodoethanol	$\mathbf{I}CH_2CH_2\mathbf{OH}$	1.5713 (54)	2.1968 (54)[e]	176.5 (54)	
38	2-phenylethanol	$\mathbf{HO}CH_2CH_2\mathbf{C_6H_5}$	1.5325 (54)	1.0202 (54)[e]	218.2 (54)	
39	2-nitroethanol	$\mathbf{NO_2}CH_2CH_2\mathbf{OH}$	1.4447 (51)	1.296 (51)[e]	195 (51)	
40	1-chloro-3,3-dimethylbutane	$\mathbf{Cl}CH_2CH_2\mathbf{C(CH_3)_3}$	1.4161 (54)	0.8670 (54)[e]	117.5 (51)	
41	1-bromo-3,3-dimethylbutane	$\mathbf{Br}CH_2CH_2\mathbf{C(CH_3)_3}$	1.4440 (51)	1.1556 (51)[e]	138 (51)	
42	1-hydroxy-3,3-dimethylbutane	$\mathbf{HO}CH_2CH_2\mathbf{C(CH_3)_3}$	1.4115 (51)[f]	0.8097 (51)	143 (54)	
43	propan-2-one	$\mathbf{CH_3}CH_2CH_2\mathbf{COCH_3}$	1.3908 (51)	0.7994 (51)	102.4 (51)	
44	1-phenylbutan-2-one	$\mathbf{CH_3CO}CH_2CH_2\mathbf{C_6H_5}$	1.5108 (51)	0.9849 (54)[j]	233.5 (54)	
45	3-phenylpropanenitrile	$\mathbf{CN}CH_2CH_2\mathbf{C_6H_5}$	1.5266 (54)	1.0016 (54)[e]	261 (54)	
46	3-chloropropanenitrile	$\mathbf{Cl}CH_2CH_2\mathbf{CN}$	1.4370 (51)	1.1573 (54)[e]	175.5 (54)	
47	1-chlorobutan-3-one	$\mathbf{Cl}CH_2CH_2\mathbf{COCH_3}$	1.4284 (54)[k]	1.0680 (54)[k]	120.5 (54)	
48	1-chloro-2-nitroethane	$\mathbf{Cl}CH_2CH_2\mathbf{NO_2}$	1.4500 (51)	1.343 (51)	173 (51)	
49	1-amino-2-phenylethane	$\mathbf{NH_2}CH_2CH_2\mathbf{C_6H_5}$	1.5315 (51)[f]	0.9640 (56)	194.5 (56)	
50	1-nitro-3,3,3-trifluoropropane	$\mathbf{NO_2}CH_2CH_2\mathbf{CF_3}$	1.3525 (51)	1.4259 (51)[e]	135.5 (51)	
51	methyl 3-cyanopropanoate	$\mathbf{CN}CH_2CH_2\mathbf{COOCH_3}$	1.4243 (51)	1.0792 (51)[e]	215.8 (51)[n]	
52	1-nitrobutane	$\mathbf{NO_2}CH_2CH_2\mathbf{CH_2CH_3}$	1.41019 (51)	0.9673 (51)	152.77 (51)	
53	butanedial	$\mathbf{OCH}CH_2CH_2\mathbf{CHO}$	1.4260 (51)	1.0659 (51)[e]	169.5 (54)	
54	hexane-2,5-dione	$\mathbf{CH_3CO}CH_2CH_2\mathbf{COCH_3}$	1.423 (51)	0.9740 (51)[e]	191.4 (49)	
55	3-bromopropanenitrile	$\mathbf{Br}CH_2CH_2\mathbf{CN}$	1.4800 (54)	1.6234 (51)		
56	1-nitro-2-phenylethane	$\mathbf{NO_2}CH_2CH_2\mathbf{C_6H_5}$	1.5270 (51)	1.119 (61)		
57	methyl 3-nitropropanoate	$\mathbf{NO_2}CH_2CH_2\mathbf{COOCH_3}$	1.4350 (51)	1.2486 (51)[e]		
58	1-nitro-3,3,3-trichloropropane	$\mathbf{NO_2}CH_2CH_2\mathbf{CCl_3}$	1.4899 (51)	1.5320 (61)[e]		
59	1-nitro-4-methylbutane	$\mathbf{NO_2}CH_2CH_2\mathbf{CH(CH_3)_2}$	1.4171 (51)	0.9458 (51)		
60	3,3,3-trichloropentane	$\mathbf{CH_3CH_2}CH_2CH_2\mathbf{CCl_3}$	1.4540 (51)	1.1843 (51)		
61	methyl 4-oxobutanoate	$\mathbf{OCH}CH_2CH_2\mathbf{COOCH_3}$	1.4210 (51)	1.087 (51)[e]		
62	1-chloro-2-iodoethane	$\mathbf{Cl}CH_2CH_2\mathbf{I}$	1.5615 (51)[f]		140 (54)	
63	3-hydroxypropanenitrile	$\mathbf{CN}CH_2CH_2\mathbf{OH}$	1.4240 (54)		230 (54)	
64	3-aminopropanenitrile	$\mathbf{CN}CH_2CH_2\mathbf{NH_2}$	1.4396 (56)		185 (56)	
65	methyl pentanoate	$\mathbf{CH_3CH_2}CH_2CH_2\mathbf{COOCH_3}$	1.3969 (51)		127.9 (51)	
66	3,3,3-trifluorobutanal	$\mathbf{CF_3}CH_2CH_2\mathbf{CHO}$	1.3387 (51)[f]		95.5 (51)	
67	1-nitrobutan-3-one	$\mathbf{NO_2}CH_2CH_2\mathbf{COCH_3}$	1.4392 (51)			

[a] Measured at 20 °C when not otherwise specified. [b] ρ/(g cm^{-3}), measured at 25 °C when not otherwise specified. [c] bp/°C, measured at 760 mmHg when not otherwise specified. [d] Measured at 25 °C when not otherwise specified. [e] Measured at 20 °C. [f] Measured at 25 °C. [g] Measured at 15 °C. [h] Measured at 18 °C. [i] Measured at 21 °C. [j] Measured at 22 °C. [k] Measured at 23 °C. [l] Measured at 30 °C. [m] Measured at 750 mmHg. [n] Measured at 753.5 mmHg.

Table 2. List of Training Set Compounds: Their Log(dynamic viscosity) Data and References (Numbers in Parentheses)

no.	compound	formula	log (η[a])
1	1,4-dioxane	$\mathbf{OCH_2CH_2O}\mathrm{CH_2CH2}$	0.081 (63)
2	ethanol	$\mathbf{HO}\mathrm{CH_2CH_2}\mathbf{H}$	0.033 (49)
3	ethane-1,2-diol	$\mathbf{HO}\mathrm{CH_2CH_2}\mathbf{OH}$	1.234 (62)
4	2-methoxyethanol	$\mathbf{HO}\mathrm{CH_2CH_2}\mathbf{OCH_3}$	0.189 (62)
5	1,2-dimethoxyethane	$\mathbf{CH_3O}\mathrm{CH_2CH_2}\mathbf{OCH_3}$	−0.378 (63)
6	2-chloroethanol	$\mathbf{HO}\mathrm{CH_2CH_2}\mathbf{Cl}$	0.429 (49)[b]
7	1,2-dichloroethane	$\mathbf{Cl}\mathrm{CH_2CH_2}\mathbf{Cl}$	−0.102 (63)
8	1-fluoro-1-chloroethane	$\mathbf{F}\mathrm{CH_2CH_2}\mathbf{Cl}$	−0.251 (51)[b]
9	bromoethane	$\mathbf{H}\mathrm{CH_2CH_2}\mathbf{Br}$	−0.421 (49)
10	1,2-dibromoethane	$\mathbf{Br}\mathrm{CH_2CH_2}\mathbf{Br}$	0.173 (49)[b]
11	iodoethane	$\mathbf{H}\mathrm{CH_2CH_2}\mathbf{I}$	−0.268 (49)[b]
12	phenylethane	$\mathbf{H}\mathrm{CH_2CH_2}\mathbf{C_6H_5}$	−0.196 (49)
13	nitroethane	$\mathbf{H}\mathrm{CH_2CH_2}\mathbf{NO_2}$	−0.180 (49)
14	chloropropane	$\mathbf{CH_3}\mathrm{CH_2CH_2}\mathbf{Cl}$	−0.498 (49)[b]
15	bromopropane	$\mathbf{CH_3}\mathrm{CH_2CH_2}\mathbf{Br}$	−0.338 (49)[b]
16	1-propanol	$\mathbf{CH_3}\mathrm{CH_2CH_2}\mathbf{OH}$	0.302 (49)
17	2-aminoethanol	$\mathbf{NH_2}\mathrm{CH_2CH_2}\mathbf{OH}$	1.159 (51)[b]
18	1,2-diaminoethane	$\mathbf{NH_2}\mathrm{CH_2CH_2}\mathbf{NH_2}$	0.188 (49)
19	butan-2-one	$\mathbf{CH_3CO}\mathrm{CH_2CH_2}\mathbf{H}$	−0.415 (63)
20	propanenitrile	$\mathbf{H}\mathrm{CH_2CH_2}\mathbf{CN}$	−0.410 (49)[b]
21	1-iodopropane	$\mathbf{CH_3}\mathrm{CH_2CH_2}\mathbf{I}$	−0.174 (49)[b]
22	1-nitropropane	$\mathbf{CH_3}\mathrm{CH_2CH_2}\mathbf{NO_2}$	−0.098 (49)
23	butanenitrile	$\mathbf{CH_3}\mathrm{CH_2CH_2}\mathbf{CN}$	−0.288 (49)[b]
24	1-chloro-2-methoxyethane	$\mathbf{Cl}\mathrm{CH_2CH_2}\mathbf{OCH_3}$	−0.228 (51)
25	propan-2-one	$\mathbf{CH_3}\mathrm{CH_2CH_2}\mathbf{COCH_3}$	−0.338 (51)
26	valeronitrile	$\mathbf{CN}\mathrm{CH_2CH_2}\mathbf{CH_2CH_3}$	−0.152 (51)
27	2-methyl-1-butanol	$\mathbf{HOCH_2CH(CH_3)}\mathrm{CH_2CH_2}\mathbf{H}$	0.740 (18)[c]
28	3-ethyl-3-pentanol	$\mathbf{H}\mathrm{CH_2CH_2}\mathbf{C(OH)(CH_2CH_3)_2}$	0.829 (18)[c]
29	methyl cyanoacetate	$\mathbf{CN}\mathrm{CH_2}\mathbf{COO}\mathrm{CH_2}\mathbf{H}$	0.446 (18)[c]
30	2-methylbutyric acid	$\mathbf{(CH_3)_2}\mathrm{CHCH_2}\mathbf{COOH}$	0.382 (18)[c]
31	heptanoic acid	$\mathbf{CH_3(CH_2)_2}\mathrm{CH_2CH_2}\mathbf{CH_2COOH}$	0.639 (18)[c]
32	isopropyl acetate	$\mathbf{CH_3COO}\mathrm{CH}\mathbf{(CH_3)}\mathrm{CH_2}\mathbf{H}$	−0.245 (18)[c]
33	isopropylamine	$\mathbf{H}\mathrm{CH_2CH}\mathbf{(CH_3)NH_2}$	−0.419 (18)[c]
34	propyl butyrate	$\mathbf{CH_3(CH_2)_2COO}\mathrm{CH_2CH_2}\mathbf{CH_3}$	−0.080 (18)[c]
35	1-pentanol	$\mathbf{CH_3CH_2}\mathrm{CH_2CH_2}\mathbf{CH_2OH}$	0.525 (49)
36	2-methyl-2,4-pentanediol	$\mathbf{(CH_3)_2COH}\mathrm{CH_2CH}\mathbf{OHCH_3}$	1.536 (18)[c]
37	2-ethoxyethanol	$\mathbf{HO}\mathrm{CH_2CH_2}\mathbf{OCH_2CH_3}$	0.312 (18)[c]

[a] η/cP, measured at 25 °C when not otherwise specified. [b] Calculated on values of η/cP measured at 30 °C. [c] Calculated on values of η/cP measured at 20 °C.

model is far simpler when few significant theoretical indexes are involved.

In the present study, we compare the multiple linear regression models obtained by using the heuristic and best multilinear variables selection procedures, implemented in the CODESSA software package, which are basically stepwise regression procedures aimed to improve the goodness of fit, with the PLS regression,[28,29] models generated by using the GOLPE variables selection procedure,[24] which is based on statistical design and is aimed to improve the predictive ability of the models. The performance of the different regression models has been evaluated by the standard deviation of prediction errors (SDEP) parameter[30] calculated for the compounds of both the training set (internal validation) and the test set (external validation). The internal SDEP has been calculated by cross-validation using both the "leave one out" (LOO) and the random group approaches.[30]

MATERIALS AND METHODS

Property Data. The experimental values of refractive index, n_D (at 20 °C), density, ρ (at 25 °C), boiling point, bp (at 760 mmHg), static dielectric constant, ϵ (at 25 °C), and dynamic viscosity, η (at 25 °C), have been taken from the literature and are listed in Tables 1 and 2 (training set) and in Table 3 (test set) together with the relevant references.

If the values of the properties (ϵ, n_D, ρ, η) for some compounds were not available at the chosen temperature, we referred to measures taken within a ±5 °C range; analogously, boiling temperatures of four compounds have been taken at different pressures ranging from 750 to 756 mmHg, as specified in Tables 1−3.

Calculations. Molecular geometry data (bond distances, bond angles, dihedral angles) have been taken from literature experimental data[31] when available. Otherwise, we referred to standard geometries taken from ab initio compilation.[32]

Molecular structures have been considered in the anti conformation with regard to the $-CH_2-CH_2-$ bond, unless otherwise specified in the literature.

Full geometry optimization of the molecular structures has been made at a semiempirical level (AM1 Hamiltonian) using the MOPAC (version 6.0)[33,34] software package.

Molecular Descriptors. The calculated molecular descriptors employed in this study, not given here but available on request to the authors, have been calculated by the CODESSA (version 2.14) software package.[22] They can be divided into five groups[7]: (1) constitutional descriptors; (2) topological descriptors; (3) geometric descriptors; (4) quantum-chemical descriptors; (5) "mixed" descriptors. "Mixed" descriptors reflect electronic, geometric, and topological features of the molecule; in particular, we considered the

Table 3. List of Test Set Compounds: Their Property Data and References (Numbers in Parentheses)

no.	compound	formula	n_D[a]	ρ[b]	bp[c]	ϵ[d]	log (η[e])
1	2-ethyl-1-hexanol	$\mathbf{HOCH_2CH(CH_2CH_3)}CH_2CH_2\mathbf{CH_2CH_3}$					0.991 (18)
2	1-butanol	$\mathbf{CH_3CH_2}CH_2CH_2\mathbf{OH}$	1.3992 (49)	0.8021 (49)[m]	117.726 (49)	17.1 (49)	0.356 (49)[m]
3	1-chlorobutane	$\mathbf{Cl}CH_2CH_2\mathbf{CH_2CH_3}$	1.4021 (49)	0.8864 (49)[g]	78.44 (49)	7.39 (49)	−0.393 (49)[m]
4	2,4-dimethylpentane	$\mathbf{(CH_3)_2CH}CH_2CH\mathbf{(CH_3)_2}$	1.3815 (49)	0.6683 (49)	80.5 (49)	1.914 (49)[g]	−0.443 (49)[g]
5	2-butanol	$\mathbf{CH_3CHOH}CH_2CH_2\mathbf{H}$	1.3978 (49)	0.8027 (49)	99.529 (49)	15.8 (49)	0.624 (49)[g]
6	3-methyl-1-butanol	$\mathbf{HO}CH_2CH_2\mathbf{CH(CH_3)_2}$	1.4085 (49)[f]	0.8018 (49)[m]	132 (49)	14.7 (49)	0.471 (49)[m]
7	ethyl cyanoacetate	$\mathbf{CNCH_2COO}CH_2CH_2\mathbf{H}$	1.4175 (49)	1.0564 (49)	206 (49)	26.9 (49)[g]	0.398 (49)
8	diethyl ether	$\mathbf{CH_3CH_2O}CH_2CH_2\mathbf{H}$	1.3527 (49)	0.7078 (49)	34.481 (49)	4.335 (49)[g]	−0.616 (49)[g]
9	isopropylbenzene	$\mathbf{C_6H_5}CH\mathbf{CH_3}CH_2\mathbf{H}$	1.4915 (49)	0.8575 (49)	152.393 (49)	2.38 (49)[g]	−0.131 (49)
10	1,3-propanediol	$\mathbf{HO}CH_2CH_2\mathbf{CH_2OH}$	1.4396 (49)	1.053 (49)[g]	214.22 (49)	35 (49)[g]	
11	1-pentene	$\mathbf{CH_2{=}CH_2}CH_2CH_2\mathbf{CH_3}$	1.3714 (49)	0.6359 (49)	29.97 (49)	2.1 (49)[g]	
12	butyraldehyde	$\mathbf{CH_3}CH_2CH_2\mathbf{CHO}$	1.3791 (49)	0.8016 (49)[g]	74.78 (49)	13.4 (49)[k]	
13	*n*-butylamine	$\mathbf{CH_3CH_2}CH_2CH_2\mathbf{NH_2}$	1.4009 (49)	0.7341 (49)[l]	76.2 (49)[n]	5.3 (49)[h]	
14	2-ethoxyethanol	$\mathbf{CH_3CH_2O}CH_2CH_2\mathbf{OH}$	1.4080 (54)	0.9297 (54)	135 (54)	13.38 (54)	
15	2-methyl-1-butanol	$\mathbf{HOCH_2CH(CH_3)}CH_2CH_2\mathbf{H}$	1.4092 (54)	0.815 (54)	128 (54)	15.63 (54)	
16	3-ethyl-3-pentanol	$\mathbf{H}CH_2CH_2\mathbf{C(OH)(CH_2CH_3)_2}$	1.4294 (54)	0.8407 (54)	142 (54)	3.158 (54)[g]	
17	isopropylamine	$\mathbf{HCH_2}CH\mathbf{(CH_3)NH_2}$	1.3742 (54)	0.6889 (54)	32.4 (54)	5.6268 (54)[g]	
18	methyl cyanoacetate	$\mathbf{CN}CH_2\mathbf{COO}CH_2\mathbf{H}$	1.4176 (54)	1.1225 (54)	200.5 (54)	28.8 (54)[g]	
19	propyl butyrate	$\mathbf{CH_3(CH_2)_2COO}CH_2CH_2\mathbf{CH_3}$	1.4001 (54)	0.8730 (54)	143 (54)	4.3 (54)[g]	
20	1-pentanol	$\mathbf{CH_3CH_2}CH_2CH_2\mathbf{CH_2OH}$	1.4101 (54)	0.8144 (54)	137.9 (54)	15.13 (54)	
21	2-methylbutyric acid	$\mathbf{(CH_3)_2}CHCH_2\mathbf{COOH}$	1.4051 (54)	0.9682 (55)	177 (54)		
22	heptanoic acid	$\mathbf{CH_3(CH_2)_2}CH_2CH_2\mathbf{CH_2COOH}$	1.4170 (54)	0.9181 (54)	222.2 (54)		
23	isopropyl acetate	$\mathbf{CH_3COO}CH\mathbf{(CH_3)}CH_2\mathbf{H}$	1.3773 (54)	0.8718 (54)	88.6 (54)		
24	1,4-dibromobutane	$\mathbf{BrCH_2}CH_2CH_2\mathbf{CH_2Br}$	1.5190 (54)	1.789 (54)[g]	197 (54)		
25	1,4-dichlorobutane	$\mathbf{ClCH_2}CH_2CH_2\mathbf{CH_2Cl}$	1.4542 (54)	1.1408 (54)[g]	153.9 (54)		
26	2-bromoethylacetate	$\mathbf{CH_3COO}CH_2CH_2\mathbf{Br}$	1.4570 (54)[j]	1.514 (54)[g]	162.5 (54)		
27	(2-bromo-1-hydroxyethyl)benzene	$\mathbf{C_6H_5}CH\mathbf{(OH)}CH_2\mathbf{Br}$	1.5800 (54)[f]	1.4994 (54)[g]			
28	1,4-diiodobutane	$\mathbf{ICH_2}CH_2CH_2\mathbf{CH_2I}$	1.619 (540[j]	2.349 (54)[k]			
29	2-methyl-2,4-pentanediol	$\mathbf{(CH_3)_2COH}CH_2CH\mathbf{OHCH_3}$	1.4276 (54)		197.1 (54)		
30	4-methylpentanenitrile	$\mathbf{(CH_3)_2CH}CH_2CH_2\mathbf{CN}$		0.8027 (49)	153.5 (49)[o]	15.5 (49)[i]	

[a] Measured at 20 °C when not otherwise specified. [b] ρ/g cm^{-3}, measured at 25 °C when not otherwise specified. [c] bp/°C, measured at 760 mmHg when not otherwise specified. [d] Measured at 25 °C when not otherwise specified. [e] η/cP, measured at 25 °C when not otherwise specified. [f] Measured at 15 °C. [g] Measured at 20 °C. [h] Measured at 21 °C. [i] Measured at 22 °C. [j] Measured at 25 °C. [k] Measured at 26 °C. [l] Measured at 27.9 °C. [m] Measured at 30 °C. [n] Measured at 752 mmHg. [o] Measured at 756 mmHg.

charged partial surface area (CPSA) descriptors defined by Stanton and Jurs,[35,36] which are derived from the net atomic charge distribution on the solvent-accessible surface.

The procedures for the calculation of a variety of constitutional, topological, geometric, and electrostatic descriptors are implemented in CODESSA. Within the class of constitutional descriptors, we used only the molecular and relative molecular weight, the gravitation indexes, and the total number of atoms and bonds of the molecule, thus avoiding considering the count of atom types, which may be present only for a small subset of compounds. The quantum chemical descriptors extracted from the output of the molecular orbital (MOPAC) calculations include among others, the following ones: Mulliken net atomic charges, the total dipole moment of the molecule and its components, the frontier molecular orbital energies and the relevant reactivity indexes, molecular polarizability terms, bond orders, and energy partitioning terms.[7] The quantum chemically calculated atomic net charges were used to calculate the CPSA descriptors.

The complete set of descriptors has been calculated by considering both the whole molecule and a molecular fragment (this is highlighted in boldface characters in the molecular formulas shown in Tables 1−3) constituting the variable portion of the considered molecular series. The studied compounds can all be represented by the general formula $X{-}CH_2CH_2{-}Y$; we considered the group composed by the X plus Y substituents as the variable fragment. Six of the test set compounds (2,4-dimethylpentane, isopropylbenzene, isopropylamine, 2-methylbutyric acid, isopropyl acetate, 2-methyl-2,4-pentanediol), in which one of the two ethyl carbons is disubstituted, have general formula $X{-}CH_2{-}CH = YZ$; in this case we considered the X, Y, and Z substituents as the variable fragment. The descriptors related to the variable fragments have been characterized by the prefix f.

After discarding all the descriptors bearing missing values and those with zero variance, 354 descriptors were left.

Multilinear Regressions. Multilinear regressions have been calculated using the CODESSA software package[6,22] that furnishes different tools to select the most promising QSPRs; in particular, we employed the heuristic (HEUR) and the best multilinear regression (BMLR) procedures, which are described in detail in the literatures.[8] Briefly, both procedures are based on the stepwise regression technique to obtain the best multiregression models and start from a collinearity control of the descriptors. The HEUR procedure is aimed at obtaining the best 1-to-n-parameter correlations (where n is a user-specified value): in this study we set n to 3. For a few properties, we have also considered $n = 4$ in order to have the same number of parameters for the HEUR regressions as for the regression models derived by the other procedures. The starting number of descriptors is reduced on the basis of the statistical significance (R^2, F, and t parameters are considered) of the 1-parameter correlations and of the descriptors' intercorrelations; finally, an iterative procedure is used to choose the best n-parameter correlations. However, this procedure only gives the best relationships derived from the best 2-parameter correlations (the default option in CODESSA is to consider the three best 2-parameter correlations: we use instead the 10 best ones) and not, for a given set of descriptors, the best relationships "altogether".

Table 4. Statistical Parameters of Multilinear Regression and PLS Models with Selected Variables for the Refractive Index (n_D) Data Set (Tables 1 and 3)[a]

				training set (67 compounds) (SDEV = 0.0559)						test set (SDEV = 0.0593)	
variables selection procedure	no. of variables	selected variables	no. of PLS components (GOLPE)	R^2	SDEC	R^2_{CV} (LOO)	SDEP$_i$	R^2_{CV} (10 groups, 30 cycles)	SDEP$_i$	SDEP$_e$ (28 compds)[b]	SDEP$_e$ (29 compds)
heuristic, BMLR	2	HOMO-1 energy, relative molecular weight		0.6739	0.0332	0.6432	0.0335	0.6401	0.0335	0.0219	0.0253
heuristic	3	HOMO-1 energy, av valency of a H atom, f-FNSA-3 (PNSA-3/TFSA)		0.7953	0.0265	0.7609	0.0273	0.7604	0.0273	0.0232	0.0314
BMLR	3	final heat of formation, molecular weight, HACA-1/TMSA		0.7893	0.0265	0.7560	0.0276	0.7552	0.0277	0.0209	0.0247
heuristic, BMLR	4	final heat of formation, molecular weight, HOMO energy, av valency of a H atom		0.8828	0.0192	0.8486	0.0218	0.8465	0.0219	0.0253	0.0250
GOLPE	20		4	0.9501	0.0125	0.9197	0.0159	0.9195	0.0159	0.0152	0.0180

[a] The limit of chance correlation is given by SDEV = $(SSY/N)^{1/2}$. [b] Omitted 1,4-diiodobutane (compound 28, Table 3).

Table 5. Statistical Parameters of Multilinear Regression and PLS Models with Selected Variables for the Density (ρ) Data Set (Tables 1 and 3)[a]

				training set (61 compounds) (SDEV = 0.3479)						test set (SDEV = 0.3712)
variables selection procedure	no. of variables	selected variables	no. of PLS components (GOLPE)	R^2	SDEC	R^2_{CV} (LOO)	SDEP$_i$	R^2_{CV} (10 groups, 30 cycles)	SDEP$_i$	SDEP$_e$ (28 compds)
heuristic	1	relative molecular weight		0.9290	0.0943	0.9229	0.0966	0.9225	0.0972	0.0899
heuristic	2	relative molecular weight, f-Min net atomic charge		0.9599	0.0714	0.9539	0.0747	0.9533	0.0759	0.0716
GOLPE	4	relative molecular weight, molecular weight, Kier and Hall index (order 1), Tot Mol 1-center 3-n attr/ no. of atoms	3	0.9481	0.0793	0.9370	0.0873	0.9361	0.0879	0.0926

[a] The limit of chance correlation is given by SDEV = $(SSY/N)^{1/2}$.

Table 6. Statistical Parameters of Multilinear Regression and PLS Models with Selected Variables for the Boiling Point Data Set (Tables 1 and 3)[a]

				training set (59 compounds) (SDEV = 53.24)						test set (SDEV = 55.45)
variables selection procedure	no. of variables	selected variables	no. of PLS components (GOLPE)	R^2	SDEC	R^2_{CV} (LOO)	SDEP$_i$	R^2_{CV} (10 groups, 30 cycles)	SDEP$_i$	SDEP$_e$ (27 compds)
heuristic	2	SQRC (gravitation index all bonds) HA-dependent HDCA-2/SQRT(TMSA)		0.7811	25.57	0.7561	26.29	0.7548	26.36	26.27
heuristic	3	SQRC (gravitation index all bonds), HA-dependent HDCA-2/TMSA, final heat of formation		0.8670	20.17	0.8408	21.24	0.8387	21.38	23.66
heuristic	4	HA-dependent HDCA-2/TMSA, SQRT (gravitation Index all bonds), final heat of formation, HOMO energy		0.9131	16.40	0.8863	18.02	0.8818	18.30	16.23
BMLR	2	HA-dependent HDCA-2, SQRC (gravitation index all bonds)		0.7791	25.69	0.7539	26.41	0.7524	26.49	28.83
BMLR	3	bonding information content (order 0), f-FHDCA fractional HDCA (HDCA/TMSA), final heat of formation		0.8663	20.16	0.8461	20.88	0.8451	20.95	28.82
BMLR	4	HA-dependent HDCA-2, SQRT (gravitation index all bonds), final heat of formation, HOMO energy		0.9074	16.94	0.8794	18.48	0.8796	18.47	19.46
GOLPE	20		4	0.9315	13.93	0.8947	17.27	0.8949	17.26	18.89

[a] The limit of chance correlation is given by SDEV = $(SSY/N)^{1/2}$.

In other words, a hypothetical *n*-parameter good correlation could not be retrieved if the whole of the 2-parameter equations obtained using 2 of the *n* parameters does not lead to satisfactory R^2 (or *F*) values. On the other hand, the BMLR procedure allows one to find the best *n*-parameter multilinear regressions altogether; iteratively, all the best N_c ($\leq$400) $n - 1$ parameter correlations are considered as candidates to perform *n*-parameter correlations until no further improvement (estimated by the *F* parameter) is achieved. Anyway, since this procedure is computationally intensive, it is not

Table 7. Statistical Parameters of Multilinear Regression and PLS Models with Selected Variables for the Dielectric Constant (ϵ) Data Set (Tables 1 and 3)[a]

				training set (23 compounds) (SDEV = 9.840)						test set (SDEV = 9.253)
variables selection procedure	no. of variables	selected variables	no. of PLS components (GOLPE)	R^2	SDEC	R^2_{CV} (LOO)	$SDEP_i$	R^2_{CV} (5 groups, 30 cycles)	$SDEP_i$	$SDEP_e$ (20 compds)
heuristic	2	f-HASA-1/TFSA, max net atomic charge		0.9120	3.130	0.8721	3.521	0.8713	3.531	5.107
heuristic	3	f-HASA-1/TFSA, RPCG relative positive charge, f-av bond inf cont (ord 0)		0.9505	2.409	0.9239	2.714	0.9190	2.800	3.999
BMLR	3	HOMO energy, f-HASA-2-/TFSA max net atomic charge		0.9564	2.262	0.9162	2.848	0.9160	2.852	4.650
GOLPE	15		3	0.9744	1.576	0.9329	2.550	0.9245	2.704	3.213

[a] The limit of chance correlation is given by SDEV = $(SSY/N)^{1/2}$.

Table 8. Statistical Parameters of Multilinear Regression and PLS Models with Selected Variables for the Log(dynamic viscosity), Log (η), Data Set (Tables 2 and 3)[a]

				training set (37 compounds) (SDEV = 0.504)						test set (SDEV = 0.522)
variables selection procedure	no. of variables	selected variables	no. of PLS components (GOLPE)	R^2	SDEC	R^2_{CV} (LOO)	$SDEP_i$	R^2_{CV} (7 groups, 30 cycles)	$SDEP_i$	$SDEP_e$ (9 compds)
heuristic, BMLR	2	f-HA-dependent HDCA-2/SQRT(TMSA), SQRC (grav ind all pairs)		0.8086	0.2206	0.7675	0.2433	0.7610	0.2465	0.1995
heuristic	3	f-HA-dependent HDCA-2/SQRT(TMSA), SQRC (grav ind all pairs), (1/2)X BETA polarizability (DIP)		0.8491	0.1959	0.8145	0.2173	0.8079	0.2210	0.1895
heuristic	4	f-HA-dependent HDCA-2/SQRT(TMSA), SQRC (grav ind all pairs), f-information content (order 0), HOMO–LUMO energy gap		0.8993	0.1601	0.8420	0.2004	0.8407	0.2012	0.2328
BMLR	3	f-HA-dependent HDCA-2/SQRT(TMSA), SQRC (grav ind all pairs), max e–e repulsion for a C–H bond		0.8595	0.1889	0.8212	0.2134	0.8162	0.2162	0.2210
BMLR	4	f-HA-dependent HDCA-2/SQRT(TMSA), SQRC (grav ind all pairs), max e–e repulsion for a C–H bond, min valency of a C atom		0.8865	0.1698	0.8425	0.2004	0.8366	0.2038	0.2302
GOLPE	16		4	0.9497	0.1131	0.9031	0.1570	0.8948	0.1635	0.2911

[a] The limit of chance correlation is given by SDEV = $(SSY/N)^{1/2}$.

possible to use large sets of descriptors (≥100), so we employed in the BMLR only the descriptors previously selected, for each experimental property, by the HEUR and the GOLPE (see next paragraph) procedures.

GOLPE Multivariate Analysis. The PLS regression models with the highest predictive capability have been derived by using the GOLPE variable selection procedure. As to the computational aspect of GOLPE, we refer to the original articles.[23,24] Briefly, GOLPE selects the best combination of variables through the following steps: (a) The combinations of variables are established according to a fractional factorial design (FFD),[37] where each one of the two levels (1, −1) corresponds to the presence and the absence of the variable, respectively. A design matrix is obtained with as many columns as variables and as many rows as combination of variables to be tested. (b) For each combination of variables, the prediction ability of the corresponding PLS model (where only the "plus" variables are included and regressed against the Y property) is evaluated by means of standard deviation of error of predictions (SDEP)[30] values. (c) The calculated SDEP values for each combination of variables are collected in a response vector and used as Y variable in another PLS model where the X-block is constituted by the design matrix. (d) The optimal model can thus be derived using only those variables proved to be significant for lowering the SDEP values. To prevent the risk of selecting as significant a variable that is actually not, a number of dummy variables can be introduced anywhere in the design matrix. The introduction of these dummy variables allows the comparison between the effect of a true variable and the average effect of the dummies. (e) On the basis of the effect on SDEP, the variables are classified as dummies, as variables with surely positive (which will be fixed) or negative effect (which will be excluded) on model predictivity, and as variables with uncertain effect. The procedure is repeated iteratively until variables are neither fixed nor excluded. According to the authors' suggestions,[23,38] the design matrix was formed with a 2:1 ratio of combinations/variables number and a 2:1 ratio of true/dummy variables, respectively. The calculation of SDEP during the selection steps was performed by the LOO procedure.

Before carrying out the statistical analysis, the distribution of the X variables was checked and two-level variables or variables showing strong clustering of objects were not included in the analysis. All variables were autoscaled to unit variance. Before running the FFD selection procedure, we employed a fast preselection technique based on D-OPTIMAL design[39] (also implemented in the GOLPE package) to obtain a reduction to about 30% of the number

Chart 1

NAME	DEFINITION	REF.
Relative molecular weight	Molecular weight divided by the number of atoms.	
	Geometrical descriptors	
Inertia A (B, C)	Principal moments of inertia of a molecule I_A, I_B and I_C, calculated by MOPAC.	
YZ Shadow **XY Shadow / XY Rectangle**	When the molecule is oriented in the space along the axes of inertia (X coordinate is along the main axis of inertia and so on) XY, YZ and XZ Shad. are the areas of the shadows of the molecule as projected on the XY, YZ and XZ planes. XY Shad. / XY Rect. is the normalised shadow area calculated as the ratios XY Shad. $/(X_{max}Y_{max})$, where X_{max}, Y_{max} are the maximum dimensions of the molecule along the corresponding axes.	46
SQRC (Gravitation index all bonds) **SQRC (Gravitation index all pairs)** **SQRT (Gravitation index all bonds)** **SQRT (Gravitation index all pairs)**	Square (SQRT) and cube (SQRC) roots of the gravitation index (G) for all pairs of atoms or for all bonded pairs of atoms: $G = \sum_{i<j}^{N} \frac{m_i m_j}{r_{ij}^2}$ where m_i and m_j are the atomic weights of atoms i and j, r_{ij} is the interatomic distance and N is equal to the number of atoms if the gravitation index is calculated over all pairs, or it is equal to the number of bonds if the gravitation index is calculated over all bonded pairs of atoms.	
	CPSA descriptors	
RPCG, RNCG	Relative positive and relative negative, charge, defined as the ratio between the most positively (negatively) charged atom and the sum of the total positive (negative) charges	35
RPCS, RNCS	Relative positive and relative negative charged surface area, defined by the product of the solvent accessible surface area of the most positive (negative) atom by RPCG (RNCG).	35
TFSA	Total solvent-accessible surface area of the molecular variable fragment.	35
TMSA	Total solvent-accessible molecular surface area.	35
PNSA-3	Atomic charge weighted partial negative surface area : PNSA-3 = Σ $(-SA_i)$ (q_i^-) where $-SA_i$ is the contribution of the i-th atom, having a negative partial atomic charge equal to q_i^-, to the total molecular solvent-accessible surface area	35
HASA-1	Hydrogen acceptors surface area: HASA-1 = $\sum_i SA_i$ where SA_i is the contribution of the i-th atom, being a hydrogen bonding acceptor atom, to the total molecular solvent-accessible surface area.	36
HASA-2	HASA-2 = $\sum_i (SA_i)^{1/2}$ where SA_i is the contribution of the i-th atom, being a hydrogen bonding acceptor atom, to the total molecular solvent-accessible surface area.	36
HACA-1	Hydrogen acceptors charged surface area: HACA = Σ SA_i q_i, where the sum is over all the atoms (with charge q_i) that are hydrogen bonding acceptors.	36
HDSA	HDSA = Σ SA_i where the sum is over all the atoms that are bonded to an oxygen atom of an OH group.	36
HDCA	HDCA = Σ SA_i q_i, where the sum is over all the atoms (with charge q_i) that are bonded to an oxygen atom of an OH group.	36
HA dependent HDCA-2	HA dependent hydrogen donors charged surface area: HA dependent HDCA-2 = $\sum_i q_i (SA_i)^{1/2}$ where the sum is over all the atoms bonded to an oxygen, and a nitrogen atom or to an alpha carbon with respect to a cyanide or a carbonyl group.	36
	Topological descriptors	
Randic and **Kier & Hall indices (order n)**	Calculated by the general formula: ${}^{n}\chi = \sum_{i(n)} (\delta_{i,1} \ldots \delta_{i,n+1})^{1/2}$ where δ_i and δ_j $(i \neq j)$ correspond to the coordination numbers of atoms (Randic index) or to the values of the atomic connectivity (Kier & Hall).	41
Kier shape index (order 1-3)	The shape of molecule depends on the number of skeletal atoms, the molecular branching and the special parameter a_i which is calculated as the ratio of the atomic radius (r_i) and the radius of the carbon atom in the sp3 hybridisation state (r_0): ${}^{1}\kappa = (N_{SA}+\alpha)(N_{SA}+\alpha-1)^2({}^{1}P+\alpha)^2$ ${}^{2}\kappa = (N_{SA}+\alpha-1)(N_{SA}+\alpha-2)^2({}^{2}P+\alpha)^2$ ${}^{3}\kappa = (N_{SA}+\alpha-1)(N_{SA}+\alpha-3)^2({}^{3}P+\alpha)^2$ if N_{SA} is odd ${}^{3}\kappa = (N_{SA}+\alpha-3)(N_{SA}+\alpha-2)^2({}^{3}P+\alpha)^2$ if N_{SA} is even where ${}^{n}P$ is the number of paths of the length n in the molecular skeleton, and α is the sum of the a_i parameters for all skeletal atoms minus 1.	42
Av. Information content (order k) **Information content (order k)** **Av. Struct. Information content (order k)** **Av. Bond Information content (order k)** **Struct. Information content (order k)** **Bond Information content (order k)**	The average information content is defined as: ${}^{k}IC_m = -\sum_i \frac{n_i}{n} \log_2 \frac{n_i}{n}$ where n_i is the number of atoms in the i-th class and n is the total number of atoms in the molecule. The division of atoms into different classes depends upon the coordination sphere taken into account. This leads to the indices of different order k. The information content (IC) is equal to the average information content multiplied by the total number of atoms. The structural (SIC) and bonding (BIC) information contents are defined as follows: ${}^{k}SIC = \frac{{}^{k}IC}{\log_2 n}$ ${}^{k}BIC = \frac{{}^{k}IC}{\log_2 q}$ where q is the number of edges in the structural graph of the molecule.	43
	Quantum-chemical descriptors	
HOMO, HOMO-1 and LUMO energy	Energy of the highest, the second highest occupied and lowest unoccupied molecular orbitals, respectively.	
Final heat of formation	ΔH_f, energy of the molecule in the thermodynamic standard scale (elements in ideal gas state at 298.15 K and 101.325 Pa).	
Tot. Mol. 1-center e-e rep.	Total molecular one-centre electron-electron repulsion energy: $E_{E-E}(tot) = \sum_{i=1}^{n} E_{E-E,i}$ where the sum is over all the n atoms in the molecule.	45
Tot. Mol. 1-center e-n attr.	Total molecular one-centre electron-nuclear attraction energy: $E_{E-N}(tot) = \sum_{i=1}^{n} E_{E-N,i}$ where the sum is over all the n atoms in the molecule.	45
Tot Mol. electr. int.	Total intramolecular electrostatic interaction energy: $E_C(tot) = 1/2 \sum_{i=1}^{n} E_{C,i}$ where the sum is over all the n atoms in the molecule.	45
Max (Min) net atomic charge	Maximum (minimum) value of the Mulliken net atomic charge of one of the atoms in the molecule.	
Avg (Min, Max) valency of a "X" atom	Average (minimum, maximum) value of the free valence of the "X" atomic species in the molecule. Free valence is defined as follows: $V_{f,A} = V_{max} - P_A$ where V_{max} is the maximum valency of the given atomic species and P_A is given by: $P_A = \sum_{B \neq A} P_{AB}$ P_{AB} representing the maximum bond order for a given pair of atomic species A and B $(A \neq B)$ in the molecule.	47
Max (Min) e-e rep. "X"-"Y" bond, Max (Min) e-n attr. "X"-"Y" bond	Extreme values of the contributions to the energy of the molecule due to electron-electron repulsion and to nucleus-electron attraction, respectively, for a generic X-Y bond.	45
Max (Min) exchange en. for a "X"-"Y" bond	Maximum (minimum) electronic exchange energy between given two atomic species (atoms X and Y) in the molecule.	45
Max σ-σ bond order	Maximum sigma-sigma bond order for a given pair of atomic species in the molecule.	47
Tot dipole	Total dipole moment of the molecule calculated by MOPAC.	
Tot. hyb. comp. mol. dip.	Total hybridisation component of the molecular dipole. It represents the component of the molecular dipole deriving from the monocentric hybridisation term (from MOPAC).	
Tot. Point-chrg mol. dip.	Total point-charge component of the molecular dipole. It represents the component of the molecular dipole deriving from the point-charge term (from MOPAC).	
1X beta polarizability and **(1/2)X beta polarizability**	Terms of the second-order polarizability (from MOPAC).	

of initial variables (descriptors) in three subsequent steps. The D-OPTIMAL design variables preselection was done on the PLS partial weights space by using the number of PLS components corresponding to the minimum SDEP value.

Predictive Capability Evaluation of the Regression Models. The predictive capability of the selected QSPR models (reported in Tables 4−8) for each property, derived either by CODESSA or by GOLPE, has been evaluated using cross-validation techniques.[29,40] For each model, the cross-validated or the prediction correlation coefficient (R^2_{CV}) has been calculated both with the leave one out and with the random groups procedures[30] (the number of groups, reported in Tables 4−8, depending on the number of compounds contained in the training set for each model). The SDEP is defined by eq 1, in analogy to the standard deviation of error of calculations (SDEC) (eq 2).

$$\mathrm{SDEP} = \sqrt{\frac{\sum (y_{\mathrm{PRED}} - y)^2}{n}} \qquad (1)$$

$$\mathrm{SDEC} = \sqrt{\frac{\sum (y_{\mathrm{CALC}} - y)^2}{n}} \qquad (2)$$

For each model we calculated two "internal" SDEP with the LOO and the random groups procedures, named $\mathrm{SDEP}_{i,\mathrm{LOO}}$ and $\mathrm{SDEP}_{i,\mathrm{Groups}}$, respectively, which are related to the internal stability of the correlation, and an "external" one, named SDEP_e, related to the capability to predict the values of the examined properties for the compounds belonging to the test set.

In order to use the facilities implemented in the GOLPE package[38] for the calculation of SDEP$_{i,LOO}$ and SDEP$_{i,Groups}$, the QSPR models derived by the HEUR and BMLR procedures, implemented in CODESSA, have been recalculated inside the GOLPE software package, deriving as many PLS components as the number of descriptors in each multilinear regression equation (in this case, the PLS model coincides with the multilinear regression).

RESULTS AND DISCUSSION

The selected best (see below) QSPR equations obtained by the different regression procedures for each physicochemical property are shown in Tables 4−8. In each table (in column order) the following are reported: (a) the procedure employed for the variable selection; (b) the number of selected descriptors and, for the heuristic and the BMLR procedures, the name of the selected descriptors (the meaning of the variable names is given in Chart 1); (c) the number of significant principal components for the correlations obtained using the GOLPE procedure; (d) the squared correlation coefficient (R^2) and the SDEC; (e) the squared cross-validated correlation coefficients (R^2_{CV}) and the "internal" (i.e., calculated on the training set's compounds) SDEP$_{i,LOO}$ and SDEP$_{i,Groups}$ (the number of groups and the number of cycles of SDEP calculation being specified); (f) the SDEP$_e$, calculated on the test set's compounds (the number of compounds in the test set being specified).

The evaluation of the prediction ability of the QSPR models is determined on comparing SDEP with the standard deviation of the y variable (SDEV = $(SSY/N)^{1/2}$, where SSY are the y sum of squares and N is the number of compounds).

The QSPR regression equations obtained by the HEUR procedure, reported in Tables 4−8, were not always the very best QSPR equations among the 10 selected by HEUR, but the best ones chosen after checking, by visual inspection of the descriptor vs property plot, the absence of strong object clustering. Actually, we have reported the following: (a) for n_D (Table 4), the second best and the sixth best HEUR regression equations for the 2- and 3-parameter models, respectively; (b) for bp (Table 5), the ninth best HEUR regression equation for the 3-parameter model; (c) for log (η) (Table 6), the eighth best HEUR regression equation for the 3-parameter model. As far as the QSPR models derived by the GOLPE procedure are concerned, the automatic preliminary control procedure for n-level variables implemented in this package allows one to avoid most of these problems. However, the GOLPE selected descriptors were also checked by us: for n_D we removed 3 of the 23 selected descriptors, for ϵ we removed 3 of the 18 selected descriptors, and for log (η) we removed 9 of the 25 selected descriptors. In all cases, the values of the correlation coefficients and SDEP did not vary significantly. The discarded descriptors are in general descriptors defined on a single atom or bond.

Refractive Index. The best correlation for the refractive index has been obtained using the GOLPE procedure (R^2 = 0.9501, SDEP$_{i,LOO}$ = 0.0159, and SDEP$_e$ = 0.0180) as shown in Table 4 and Figure 1, where the calculated vs predicted property values are plotted for both the training and the test set compounds. There is a general improvement in the SDEP$_e$ values when the 1,4-diiodobutane (compound 28, Table 3) is omitted: in particular, the SDEP$_e$ value of the GOLPE model lowers to 0.0152. In Figure 2 the PLS pseudoregression coefficients (The PLS loadings and partial weights can be used to reformulate the dependent variable, y, as in a MLR regression equation, $y = BX$. These pseudoregression coefficients are identical to the MLR regression coefficients if the number of PLS components equals the number of variables in X. Otherwise, they are not independent of one another and are used only for interpretative purpose, i.e., to establish which are the most significant variables in the PLS model.) of the 20 selected descriptors for the 4-component PLS model are shown. The heterogeneity of these 20 descriptors, which belong to the topological, constitutional, geometric, and quantochemical families of descriptors, demonstrates the need of considering a high number of structural features in order to adequately describe a complex experimental property such as the refractive index: the multiplicity of underlying factors highlights how this phenomenon is based on complex interactions between matter and electromagnetic radiation.

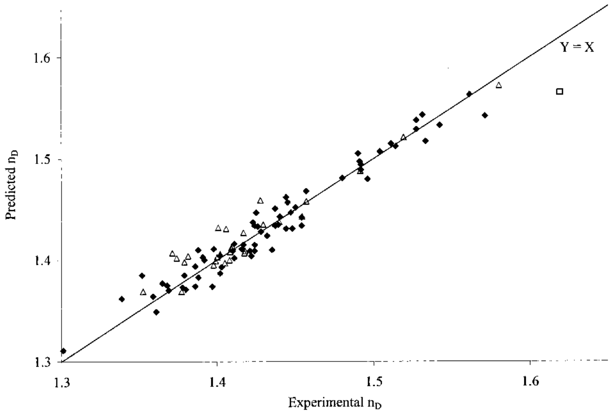

Figure 1. Plot of predicted versus experimental refractive index (n_D) values for the training set (◆) and test set compounds (△). The 1,4-diiodobutane, omitted from the correlation, is indicated by the X with vertical line. The predicted n_D values have been calculated by the GOLPE regression model, R^2 = 0.9501 in Table 4.

In agreement with what has been stated above, the multilinear regression procedures (HEUR and BMLR) failed to give correlations as satisfactory as that obtained by using the GOLPE procedure: it is noteworthy that, passing from 2 to 4 parameters, while the R^2 and SDEP$_{i,LOO}$ values improve, the SDEP$_e$ value does not.

In a recent paper[19] Katritzky et al. reported a QSPR model also derived by the CODESSA software package for the n_D of a set of 125 diverse organic compounds. R^2 = 0.945 and SDEC = 0.0155 values were computed which are similar to those of our best QSPR model.

Density. An excellent 1-parameter correlation between density and the relative molecular weight (R^2 = 0.9290, SDEP$_{i,LOO}$ = 0.0966, SDEP$_e$ = 0.0899, Table 5 and Figure 3) has been found. Although it is well known that molecular weight and density are correlated with each other within an homologous series of compounds, taking into account the structural heterogeneity of the training set's structures (Table 1), the regression model obtained appears less trivial.

Introducing a further parameter, i.e., the minimum value of the net atomic charge for the variable molecular fragment (f-Min net atomic charge), both the squared correlation coefficient (R^2 = 0.9599) and the predictive capability of

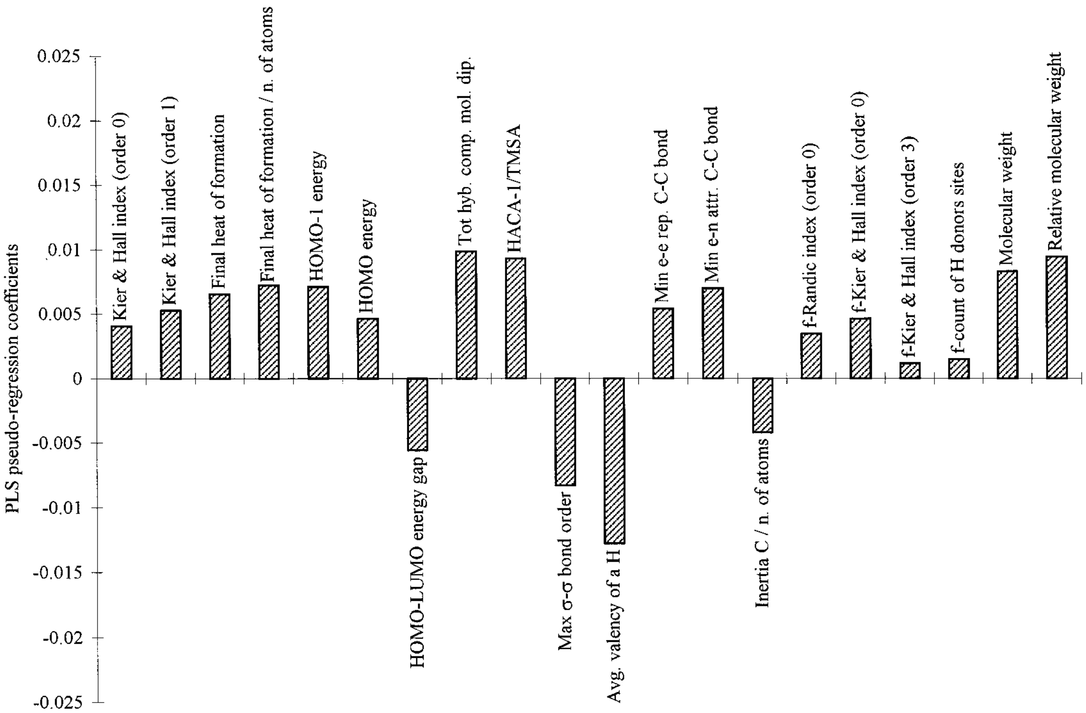

Figure 2. PLS pseudoregression coefficients of the selected variables for the refractive index, n_D (GOLPE regression model, $R^2 = 0.9501$ in Table 4).

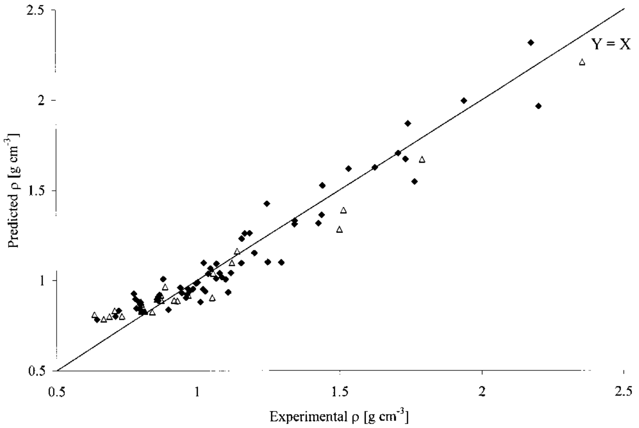

Figure 3. Plot of predicted versus experimental density (ρ (g cm^{-3})) values for the training set (◆) and test set compounds (△). The predicted ρ values have been calculated by the heuristic regression model, $R^2 = 0.9290$ in Table 5.

the regression model (SDEP$_i$ = 0.0747 and SDEP$_e$ = 0.0716) increase. This descriptor codifies to some extent the intensity of the polar interactions: the increase of its value can be reasonably related to an increase in density.

The GOLPE procedure also leads to satisfactory results; considering that just one parameter is enough to correctly predict the density values, the correlation of this experimental property represents a borderline case for the application of multivariate analysis. Therefore, it should be stressed that the GOLPE procedure, even if furnishing a redundant model, provides a really satisfactory "external" predictive capability (SDEP$_e$ = 0.0553). Furthermore, the PLS pseudoregression coefficient of the relative molecular weight presents the highest value among those of the four selected descriptors (Figure 4), confirming the convergence between this approach and the multilinear regression methods.

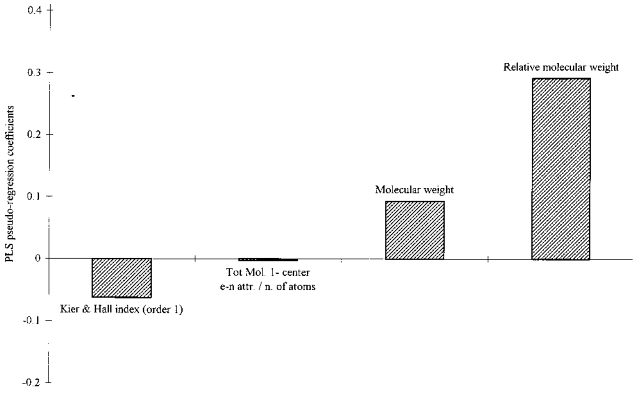

Figure 4. PLS pseudoregression coefficients of the selected variables for the density, ρ (g cm^{-3}), (GOLPE regression model, $R^2 = 0.9481$ in Table 5).

Boiling Temperature. The best QSPR model for this property has been obtained by the GOLPE procedure ($R^2 = 0.9315$, SDEP$_{i,LOO}$ = 17.27, and SDEP$_e$ = 18.89, Table 6 and Figure 5). The PLS pseudoregression coefficients for the 20 selected variables are reported in Figure 6; the involved descriptors are topological indexes that quantify both size and shape (degree of branching) of the molecule,[41−43] quantum chemical descriptors derived from the total molecular energy (final heat of formation, total molecular electrostatic interaction/number of atoms, and total 1-center electron−electron repulsion and electron−nucleus attraction) which depend on the molecular size, the molecular weight, and descriptors that quantify the hydrogen-bonding effects (HASA-1, f-HDSA, etc.).[8,36] These descriptors, though not comprising, except for the final heat of formation, those selected by the HEUR and by the BMLR procedures, have similar physical meaning. Furthermore, they have similar physical meaning with respect to those previously employed in the literature[8] for predicting the boiling temperatures of different sets of organic compounds.

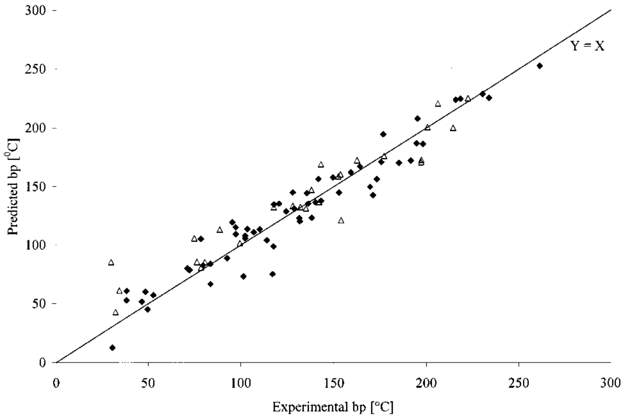

Figure 5. Plot of predicted versus experimental normal boiling point (bp (°C)) values for the training set (◆) and test set compounds (△). The predicted bp values have been calculated by the GOLPE regression model, $R^2 = 0.9315$ in Table 6.

It is worth noting that the BMLR regression models have higher $SDEP_e$ values compared to the heuristic regressions with the same number of parameters. Unexpectedly, the HEUR 4-parameter model works better than the corresponding BMLR 4-parameter model. Furthermore, it shows a better external predictive ability ($SDEP_e = 16.23$) with respect to the GOLPE model: however, the $SDEP_e$ value from GOLPE is strongly affected by the deviation of 1-pentene (compound 11, Table 3); by excluding this compound in the analysis, the $SDEP_e$ value lowers from 18.89 to 15.86. Analogously, the $SDEP_e$ value of the BMLR 3-parameter model by excluding the 1-pentene lowers from 28.82 to 24.98, which is more similar to the $SDEP_{i,Groups}$ value; a poor estimation for 1-pentene has to be expected since compounds containing a double carbon−carbon bond are not present in the training set of this property.

Our results are also consistent with a recent study of Karelson, Katritzky, et al.,[8] where CODESSA has been employed to derive QSPR models for the boiling temperatures of 298 diverse organic compounds. Their best 2-parameter regression ($R^2 = 0.9544$, SDEC = 16.15) was obtained with the SQRC (Grav Ind all bonds) and the HA-dependent HDCA-2/TMSA descriptors[36] (see Chart 1 for the definition of descriptors) which, consistently, are both present in our multilinear regressions. When this 2-parameter regression equation was used to predict the boiling temperature of our 59 training set compounds, we obtained a $SDEP_e$ value of 29.45, which is similar to the $SDEP_{i,LOO}$ value of our 2-parameter HEUR regression (Table 6) and is higher than the $SDEP_{i,LOO}$ and $SDEP_e$ values of the other QSPR regressions reported in Table 6.

Recently, Katritzky et al.[9] extended their QSPR approach to a set of 584 diverse organic compounds representative of all major classes of organic compounds containing C, H, O, N, S, F, Cl, Br, and I. The best correlations were obtained by a 6- and an 8-parameter model, with $R^2 = 0.946$, SDEC = 18.9 for the former and $R^2 = 0.9645$, SDEC = 15.5, $SDEP_i = 14.6$, and $SDEP_e = 9.68$ for the latter model, respectively. The descriptors employed in the 4-parameter regression equation are two hydrogen-bonding indexes, two size-dependent descriptors, and two variables accounting for the number of F atoms and CN groups, respectively; for the 6-parameter regression equation, two CPSA descriptors taking into account only H or Cl atoms were added. Many, though not all, of our training and test set compounds are also included in the broad calibration set reported in ref 9; however, the use of the same 6-parameter model to correlate the boiling temperature of our training set does not give improved results with respect to the models reported in Table 6.

Dielectric Constant. As it can be seen in Table 7 and Figure 7, the best regression model for the dielectric constant was obtained using the GOLPE procedure ($R^2 = 0.9744$).

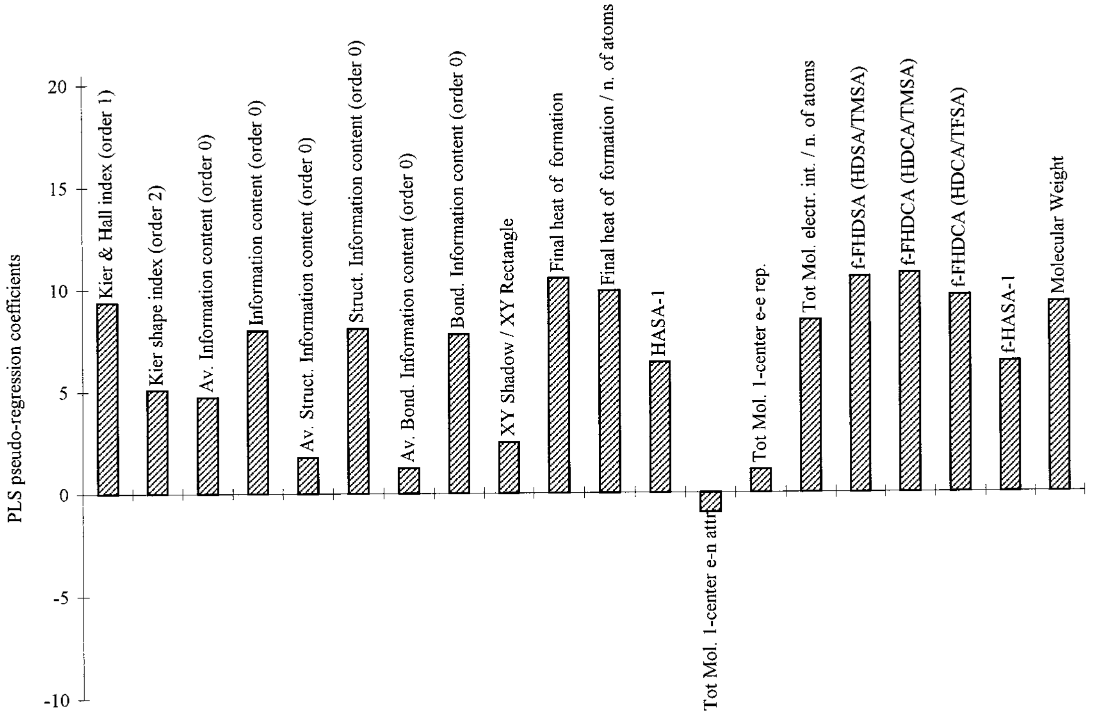

Figure 6. PLS pseudoregression coefficients of the selected variables for normal boiling point, bp (°C) (GOLPE regression model, $R^2 = 0.9315$ in Table 6).

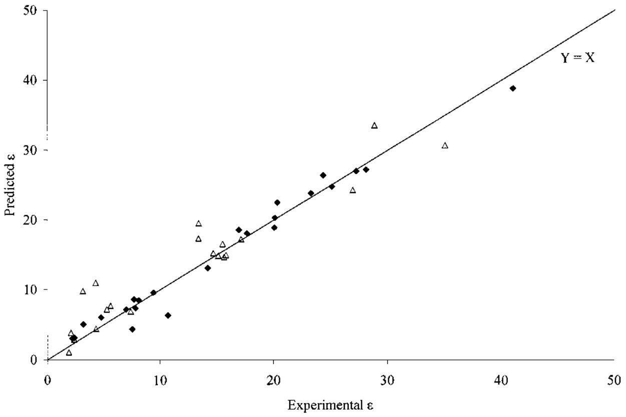

Figure 7. Plot of predicted versus experimental dielectric constant (ϵ) values for the training set (◆) and test set compounds (△). The predicted ϵ values have been calculated by the GOLPE regression model, $R^2 = 0.9744$ in Table 7.

Although the $\mathrm{SDEP}_{\mathrm{i,LOO}}$ values for the 3-parameter multilinear regressions are similar to those of the GOLPE model, the latter procedure leads to a significantly better value of $\mathrm{SDEP}_{\mathrm{e}}$.

The worst GOLPE predictions are for butyraldehyde and propyl butyrate (compounds 12 and 19 in Table 3, respectively), which has to be expected since neither aldehydes nor esters are represented in the training set of this property. Accordingly, by excluding these two compounds from the test set, the $\mathrm{SDEP}_{\mathrm{e}}$ value for the regression model calculated with the GOLPE procedure lowers from 3.213 to 2.621.

The PLS pseudoregression coefficients for the 15 descriptors selected by the GOLPE procedure are shown in Figure 8. These are descriptors related to the tendency of the molecules to act as hydrogen-bonding donors or acceptors (e.g., f-FHDCA (HDCA/TMSA) and f-HASA-2/TFSA), descriptors related to the dipole moment, topological descriptors (information content indexes[43]), charge-related descriptors (e.g., maximum net atomic charge and f-RPCS), and the HOMO energy (related to the charge-transfer tendency). The presence of molecular descriptors related to the molecular charge distribution is in agreement with a single parameter correlation ($R = 0.88$) between a charge separation index (derived from the molecular electrostatic potential) and the dielectric constant recently reported by Brink et al.[44]

In a recent study,[20] a QSPR model for the dielectric constant of a series of organic compounds has been computed using neural networks. The best QSPR relationship (the root-mean-square errors for the training set of 350 compounds and for the test set of 50 compounds are 3.77 and 2.33, respectively—the correlation coefficient is not given) selected by the authors to fit the dielectric constants uses 10 theoretical molecular descriptors, namely, the number of O and N atoms, an indicator variable for hydrogen-bonding capability, three CPSA descriptors, and three topological descriptors. CPSA and topological descriptors are also used in our QSPR models. The predictive ability of this model is comparable with our best QSPR model reported in Table 7.

Log(dynamic viscosity). The best QSPR model for this property has been obtained by the GOLPE procedure ($R^2 = 0.9497$, $\mathrm{SDEP}_{\mathrm{i,LOO}} = 0.1570$, and $\mathrm{SDEP}_{\mathrm{e}} = 0.2911$), as shown in Table 8 and Figure 9. The relevant descriptors, as can be seen in Figure 10, are hydrogen-bonding descriptors, topological descriptors, components of the molecular polarizability, and quantum mechanical energy terms for the C−C and C−H bonds which may be related to the conformational changes of the molecule.[45] The HEUR 2- and 3-parameter models show the lowest $\mathrm{SDEP}_{\mathrm{e}}$ values, but inspection of the errors distribution for the test set compounds reveals a tendency for the errors to increase with log (η); analogously, the BMLR 3-parameter model gives the highest errors for the test set compounds 2-ethyl-1-hexanol and 2-butanol (compounds 1 and 5, Table 3), which show the highest log (η) values. Furthermore, the $\mathrm{SDEP}_{\mathrm{e}}$ values for the BMLR

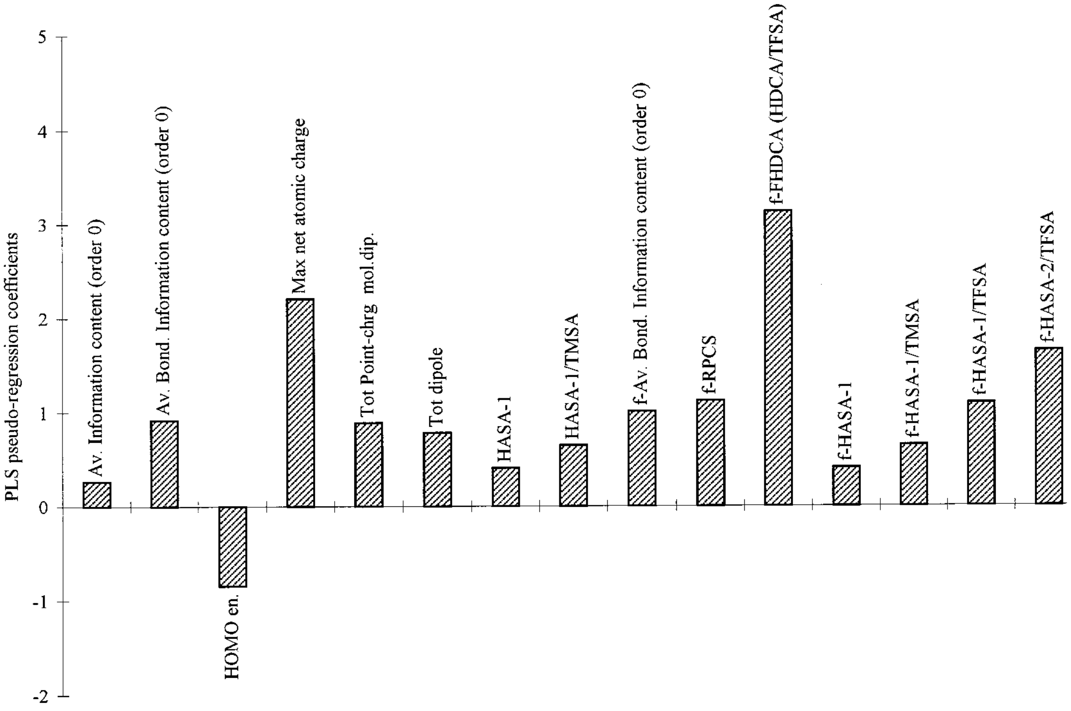

Figure 8. PLS pseudoregression coefficients of the selected variables for the dielectric constant, ϵ (GOLPE regression model, $R^2 = 0.9744$ in Table 7).

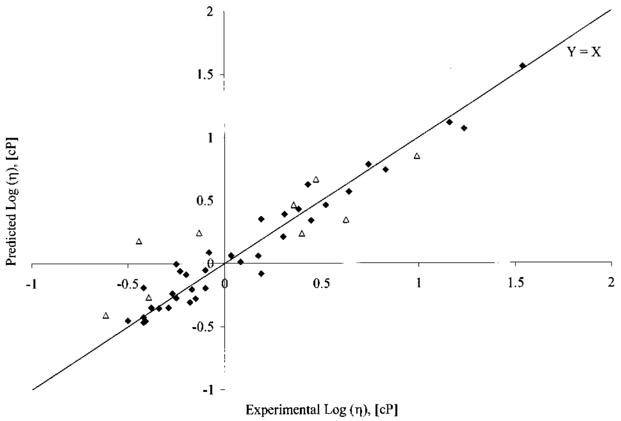

Figure 9. Plot of predicted versus experimental log(dynamic viscosity), log (η), values for the training set (◆) and test set compounds (△). The predicted log (η) values have been calculated by the GOLPE regression model, $R^2 = 0.9497$ in Table 8.

4-parameter model and for the GOLPE model improve significantly by omitting the 2,4-dimethylpentane and isopropylbenzene (compounds 4 and 9, Table 3), lowering to 0.1708 and 0.1705, respectively. The HEUR 4-parameter model gives the best $SDEP_e$ value of 0.1183 when the 2,4-dimethylpentane is omitted. A large deviation for branched compounds could be expected since they are not adequately represented in the calibration set.

Our results are consistent with a recent QSPR study of Ivanciuc et al.,[17] where CODESSA was as well employed to model the liquid viscosity of 337 diverse organic compounds. The best correlation equation reported ($R^2 = 0.8464$, SDEC = 0.371) contains five parameters: the hydrogen-bonding donor charged surface area (HDCA-2); the molecular weight; the Randic connectivity index of order 3; the maximum electrophilic reactivity index for a carbon atom; the maximum electronic population. Significantly, these descriptors are either the same or have the same meaning of those selected by our GOLPE model (Figure 10).

Another QSPR study modeling the log (η) of 237 diverse organic compounds (plus a test set of 124 additional compounds) using nine descriptors, namely, four experimental descriptors (molar refraction, critical temperature, molar magnetic susceptibility, vaporization energy) and five indicator variables (presence of alcohols/phenols, nitriles, amines, amides, and aliphatic rings including heteroatoms) and both multilinear regression and neural networks has been reported.[18] The best reported MLR and neural networks models give different degree of fit, the R^2 values being 0.916 and 0.958, respectively, but exhibit similar predictive capability, the $SDEP_e$ values being 0.168 and 0.161, respectively. The highest calculation errors were found for compounds containing several OH groups, thus indicating a poor parameterization for the hydrogen bond effect. The performance of our best QSPR model is quite similar with respect to the model fit; the predictive capability is lower but it becomes comparable when 2,4-dimethylpentane and isopropylbenzene are taken out from the set. Interestingly enough, four out of nine of our test set compounds contain an OH group. Among the compounds of our training set bearing more than one hydrogen-bonding group, only the 1,2-diaminoethane (compound 18, Table 2) is strongly underestimated [Δlog (η)(calc − exp) = −0.27], indicating that the theoretical descriptors employed by us adequately take into account hydrogen bond effects.

CONCLUSIONS

Good QSPR models have been obtained for all the studied properties, confirming that theoretical molecular descriptors computed on the isolated molecule are also suitable to both fit and predict physicochemical properties of molecular series in the condensed phases. Noteworthy, we did not use any

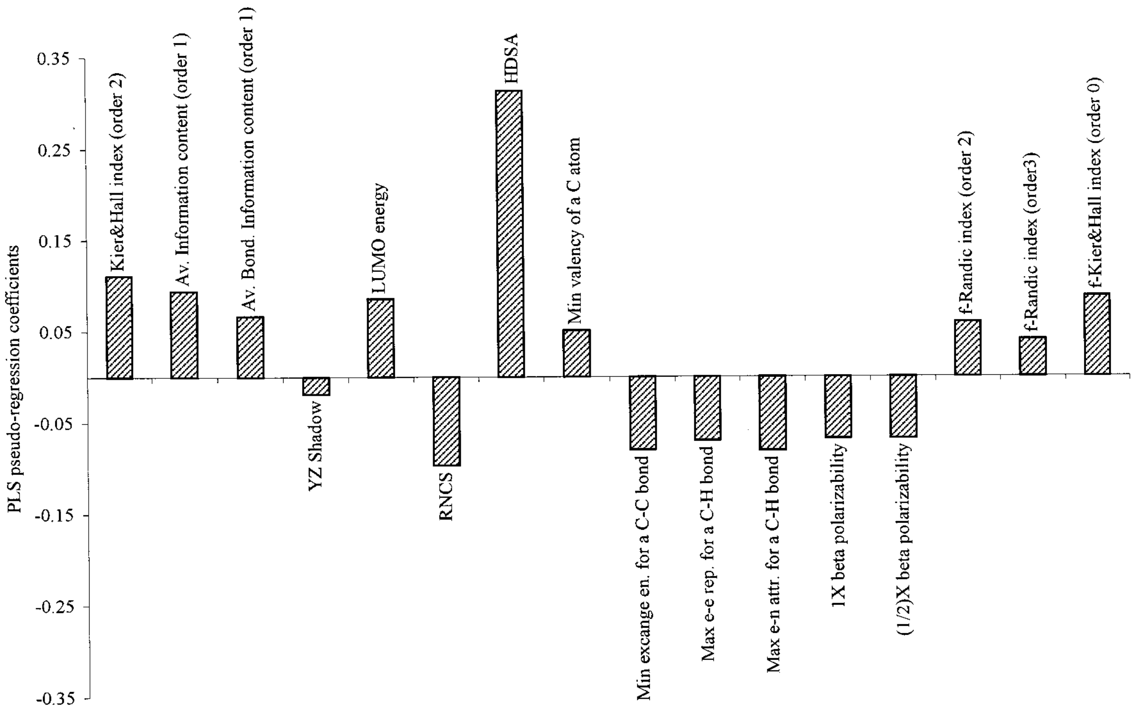

Figure 10. PLS pseudoregression coefficients of the selected variables for the log (dynamic viscosity), log (η) (GOLPE regression model, $R^2 = 0.9497$ in Table 8).

indicator variables or descriptors based on atom or atomic group count.

The comparative analysis of the HEUR and BMLR multilinear regression techniques with the multivariate GOLPE/PLS methods has shown the following: (a) the results from the GOLPE regression models are generally better; (b) the selected descriptors are often the same or have at least similar meaning; (c) at variance with expectation, the BMLR procedure does not always furnish better regressions than the HEUR procedure (see Tables 6 and 8); and (d) the difference between the $SDEP_{i,LOO}$ and the $SDEP_{i,Groups}$ values for the most part of the presented models is small, indicating a satisfactory internal stability. Furthermore, the $SDEP_i$ values are often close to the $SDEP_e$ ones, indicating the validity of this index as a validation criterion, at least when the calibration set includes the structural variability of the test set.

ACKNOWLEDGMENT

The authors thank Prof. S. Clementi and the research group for chemometrics of the University of Perugia for access to GOLPE and Prof. M. Karelson of the University of Tartu, Estonia, for access to CODESSA. Financial support from the Ministero dell'Universita' e della Ricerca Scientifica and from Consiglio Nazionale delle Ricerche (ROMA) is acknowledged.

REFERENCES AND NOTES

(1) Cramer, R. D. BC(DEF) Parameters. 2. An Empirical Structure-Based Scheme for the Prediction of Some Physical Properties. *J. Am. Chem. Soc.* **1979**, *102*, 1849−1859.

(2) Monnery, W. D.; Svreck, W. Y.; Mehrotra, A. K. Viscoity: A Critical Review of Practical Predictive and Correlative Methods. *Can. J. Chem. Eng.* **1995**, *73*, 3−40.

(3) Stein, S. E.; Brown, R. L. Estimation of Normal Boiling Points from Group Contributions. *J. Chem. Inf. Comput. Sci* **1994**, *34*, 581−587.

(4) Pouchly, J.; Quin, A.; Munk, P. Excess Volume of Mixing and Equation of State Theory. *J. Solution Chem.* **1993**, *22*, 399−418.

(5) Elbro, H. S.; Fredenslund, A.; Rasmussen, P. Group Contribution Method for the Prediction of Liquid Densities as a Function of Temperatures for Solvents, Oligomers, and Polymers. *Ind. Eng. Chem. Res.* **1991**, *30*, 2576−2593.

(6) Katritzky, A. R.; Lobanov, W. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, 279−287.

(7) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1043.

(8) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407 and references therein.

(9) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comput Sci.* **1998**, *38*, 28−41.

(10) Murugan, R.; Grendze, M. P.; Toomey, J. E., Jr.; Katritzky, A. R.; Karelson, M.; Lobanov, V. S.; Rachwal, P. Predicting Physical Properties from Molecular Structure. *CHEMTECH* **1994**, *24*, 17−34.

(11) Bodor, N.; Huang, M. J. Predicting Partition Coefficients for Isomeric Diastereoisomers of Some Tripeptide Analogs. *J. Comput. Chem.* **1991**, *12*, 1182−1186.

(12) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, W. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure−Property Relationship Treatment. *Anal. Chem.* **1994**, *66*, 1799−1807.

(13) Buydens, L.; Massart, D.; Geerlings, P. Prediction of Gas Chromatographic Retention Indexes with Topological, Physicochemical, and Quantum Chemical Parameters. *Anal. Chem.* **1983**, *55*, 738−752.

(14) Osmialowski, K.; Halkiewicz, J.; Radecki, A.; Kaliszan, R. Quantum Chemical Parameters in Correlation Analysis of Gas-Liquid Chromatographic Retention Indexes of Amines. *J. Chromatogr.* **1986**, *361*, 63−81.

(15) Grigoras, S. A Structural Approach to Calculate Physical Properties of Pure Organic Substances: The Critical Temperature, Critical Volume and Related Properties. *J. Comput. Chem.* **1990**, *11*, 493−510.

(16) Egolf, L. M.; Wessel, M. D.; Jurs, P. C. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947−956.

(17) Ivanciuc, O.; Ivanciuc, T.; Filip, P. A.; Cabrol-Bass, D. Estimation of Liquid Viscosity of Organic Compounds with a Quantitative Structure−Property Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 515−524.

(18) Suzuki, T.; Ebert, R. U.; Schüürmann, G. Development of Both Linear and Nonlinear Methods To Predict the Liquid Viscosity at 20 °C of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1122−1128.

(19) Katritzky, A. R.; Sild, S.; Karelson, M. General Quantitative Structure−Property Relationship Treatment of the Refractive Index of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 840−844.

(20) Schweitzer, R. C.; Morris, J. B. The Development of a Quantitative Structure Property Relationship (QSPR) for the Prediction of Dielectric Constants Using Neural Networks. *Anal. Chim. Acta* **1999**, *384*, 285−303.

(21) Marchetti, A.; Tassi, L. Thermophysics of Multicomponent Nonelectrolytic Solutions. In *Current Topics in Solution Chemistry;* Richard, R., Ed.; Research Trends Publisher: Trivandrum, 1997; Vol. 2, pp 63−82.

(22) *CODESSA 2.14;* Comprehensive Descriptors for Structural and Statistical Analysis; Copyright University of Florida, Gainesville, FL, 1994, p 95.

(23) Baroni, M.; Clementi, S.; Cruciani, G.; Costantino, G.; Riganelli, D.; Oberrauch, E. Predictive Ability of Regression Models. Part II: Selection of the Best Predictive PLS Model. *J. Chemom.* **1992**, *6*, 347−356.

(24) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.

(25) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.

(26) Lindgren, F.; Geladi, P.; Rännar, S.; Wold, S. Interactive Variable Selection (IVS) for PLS. Part 1: Theory and Algorithms. *J. Chemom.* **1994**, *8*, 349−363.

(27) Leardi, R.; Application of a Genetic Algorithm to Feature Selection under Full Validation Conditions and to Outlier Detection. *J. Med. Chem.* **1994**, *33*, 136−142.

(28) Geladi, P. Notes on the History and Nature of Partial Least Squares (PLS) modelling. *J. Chemom.* **1988**, *2*, 231−246.

(29) Wold, S.; Johansson, E.; Cocchi, M. PLS−Partial Least Squares Projections to Latent Structures. In *3D QSAR in Drug Design: Theory Methods and Applications;* Kubinyi, H., Ed.; ESCOM Science Publishers; Leiden, **1993**; pp 523−550.

(30) Cruciani, G.; Baroni, M.; Clementi, S.; Costantino, G.; Riganelli, D.; Skagerberg, B. Predictive Ability of Regression Models. Part I: Standard Deviation of Prediction Errors (SDEP). *J. Chemom.* **1992**, *6*, 335−346.

(31) Landolt-Bornstein *Structure Data of Free Polyatomic Molecules, II/7;* Springer-Verlag Publishers; Berlin, 1976 Landolt-Börnstein *Structure Data of Free Polyatomic Molecules, II/15;* Springer-Verlag Publishers; Berlin, 1987.

(32) Marriott, S.; Topsom, R. D. Standard Bond Lengths for Use in Ab Initio Molecular Orbital Calculations. *J. Mol. Struct (Theochem)* **1984**, *110*, 337−340.

(33) *MOPAC 6.0* QCPE 455/SGRW.

(34) Dewar, M. J. S.; Zoebisch, E.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3911.

(35) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(36) Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306−311.

(37) Box, G. E. P.; Hunter, W. G. *Statistics for Experimenters;* Wiley: New York, 1978.

(38) *GOLPE 3.0;* Multivariate Informatic Analysis; Perugia 1995 (User Manual).

(39) Mitchell, T. J. An Algorithm for the Construction of D-OPTIMAL Experimental Designs. *Technometrics* **1974**, *16*, 203−210.

(40) Wold, S. Cross Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics* **1982**, *24*, 73−77.

(41) Kier, L. B.; Hall, L. In *Molecular Connectivity in Structure-Activity Analysis;* Research Studies Press: Letchworth, England, 1986.

(42) Kier, L. B. In *Computational Chemical Graph Theory;* Rouvray, D. H., Ed.; Nova Science Publishers: New York 1990; pp 151−174.

(43) Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structure;* Wiley-Interscience: New York, 1983.

(44) Brinck, T.; Murray, J. S.; Politzer, P. Quantitative Determination of the Total Local Polarity (Charge Separation) in Molecules. *Mol. Phys.* **1992**, *76*, 609−617.

(45) Strouf, O. *Chemical Pattern Recognition*; Wiley: New York, 1986.

(46) Rohrbaugh, R. H.; Jurs, P. C. Molecular Shape and the Prediction of High-Performance Liquid Chromatographic Retention Indexes of Polycyclic Aromatic Hydrocarbons. *Anal. Chim. Acta* **1987**, *199*, 99−111.

(47) Sannigrahi, A. B. Ab Initio Molecular Orbital Calculations of Bond Index and Valency. *Adv. Quant. Chem.* **1992**, *23*, 301−351.

(48) Corradini, F.; Marchetti, A.; Tagliazucchi, M.; Tassi, L.; Tosi, G. Associating Behaviour of Mixed Liquids: Dielectric Properties of the Ethane-1,2-diol + 1,4-Dioxan Solvent System from −10 to + 80 °C. *Aust. J. Chem.* **1995**, *48*, 1193−1200.

(49) Weissberger, A.; Proskauer, E. S.; Riddick, J. A.; Toops, E. E. *Organic Solvents, Physical Properties and Methods of Purification;* Interscience Publishers: New York, 1955.

(50) Franchini, G. C.; Marchetti, A.; Tagliazucchi, M.; Tassi, L.; Tosi, G. Ethane-1,2-diol−2-methoxyethanol Solvent System. Dependence of the Relative Permittivity and Refractive Index on the Temperature and Composition of the Binary Mixture. *J. Chem. Soc., Faraday Trans.* **1991**, *87*, 2583−2588.

(51) Beilstein *Handbook of Organic Chemistry;* Beilstein Informationssysteme GmbH; Frankfurt am Main, Germany, 19xx.

(52) Corradini, F.; Marchetti, A.; Tagliazucchi, M.; Tassi, L. Static Dielectric Constants of 1,2-dichloroethane + 1,2-dimethoxyethane Binary Mixtures. *Ann. Chim. (Rome)* **1995**, *85*, 531−541.

(53) Corradini, F.; Marchetti, A.; Tagliazucchi, M.; Tassi, L.; Varini, A. Dielectric Characterization of Binary Solvents Containing 1,2-dichloroethane and 2-chloroethanol. *Bull. Chem. Soc. Jpn.* **1995**, *68*, 2187−2191.

(54) *Handbook of Chemistry and Physics, 75th ed.;* Lide, D. R., Ed.; CRC Press: Boca Raton, FL, 1995, and preceeding editions.

(55) *Physical Properties of Chemical Compounds;* Dreisbach, R. R., Ed.; Am. Chem. Soc.: Washington, DC, 1961.

(56) *The Merk Index*, 9th ed.; Windholz, M., Ed.; Merk & Co. Inc. Publishers: Rahway, NJ, 1976.

(57) Corradini, F.; Marcheselli, L.; Tassi, L.; Tosi, G. Kinematic Viscosities of 1,2-ethanediol/1,4-dioxane Binary Mixtures from −10 to +80 °C. *Bull. Chem. Soc. Jpn.* **1993**, *66*, 1886−1891.

(58) Franchini, G. C.; Marchetti, A.; Preti, C.; Tassi, L. Volumetric Behaviour of the Ethane-1,2-diol + 1,2-dimethoxyethane Binary Solvent System. *Ann. Chim. (Rome)* **1996**, *86*, 357−367.

(59) Corradini, F.; Marchetti, A.; Tagliazucchi, M.; Tassi, L.; Tosi, G. Volumetric Behaviour of 2-methoxyethanol + 1,2-dimethoxyethane Binary Mixtures. *Aust. J. Chem.* **1994**, *47*, 415−422.

(60) Corradini, F.; Marchetti, A.; Tagliazucchi, M.; Tassi, L. Static Dielectric Constants of 1,2-dichloroethane + 2-methoxyethanol + 1,2-dimethoxyethane Ternary Liquid Mixtures from −10 to 80 °C. *Fluid Phase Equilib.* **1996**, *124*, 209−220.

(61) Calculated from refs 51 and 54.

(62) Corradini, F.; Franchini, G. C.; Marchetti, A.; Tagliazucchi, M.; Tassi, L.; Tosi, G. Viscosities of 1,2-ethanediol−2-methoxyethanol Solvent Mixtures at Various Temperatures. *J. Solution Chem.* **1993**, *22*, 1019−1028.

(63) Calculated from the corresponding density and kinematic viscosity data.[57]

CI9903298