Estimation of the Liquid Viscosity of Organic Compounds with a Quantitative Structure−Property Model

Ovidiu Ivanciuc,[†,⊥] Teodora Ivanciuc,[†] Petru A. Filip,[‡] and Daniel Cabrol-Bass*[,§]

Department of Organic Chemistry, Faculty of Chemical Technology, University "Politehnica" of Bucharest, Oficiul 12 CP 243, 77206 Bucharest, Romania, Institute of Organic Chemistry, Romanian Academy, Splaiul Independentei 202B, 71141 Bucharest, Romania, and GRECFO−LARTIC, University of Nice−Sophia Antipolis, Parc Valrose, 06108 Nice Cedex, France

Received June 22, 1998

A QSPR (quantitative structure−property relationship) model for the estimation of the liquid viscosity of a large variety of organic compounds was established using the CODESSA (comprehensive descriptors for structural and statistical analysis) approach. The final model was developed with a calibration set containing 337 compounds. The multilinear regression (MLR) equation that relates ln η to five molecular descriptors has $s = 0.37$ and $r = 0.920$. The five theoretical parameters used in the QSPR model are molecular weight, Randić connectivity index of order 3, hydrogen-donor charged surface area HDCA-2 (electrostatic), maximum electrophilic reactivity index for a carbon atom, and maximum atomic orbital electronic population. The predictive ability of the MLR model was tested by the leave-20%-out cross-validation method, showing that the QSPR model is stable and can be used to obtain good predictions for compounds that were not used in the model calibration. The cross-validation statistical indices show a small decrease when compared with those obtained in the calibration phase ($s_{cv} = 0.38$, $r_{cv} = 0.917$). The QSPR models developed with CODESSA allow accurate computation of the liquid viscosity of organic compounds using simple constitutional descriptors and quantum indices which can be computed with standard quantum chemistry packages.

INTRODUCTION

The liquid viscosity of organic compounds is an important property for the simulation of the processes in chemical and petroleum industries. With the increased need of reliable data for optimization of the industrial processes, it is important to develop QSPR (quantitative structure−property relationship) models for the estimation of the viscosity for compounds not synthesized or with an unknown viscosity.[1] Numerous QSPR models for calculating the liquid viscosities have been proposed using various numerical descriptors of the chemical structure.

Suzuki and co-workers[2,3] proposed a quantitative property−property relationship (QPPR) model in which the liquid viscosity is computed on the basis of a set of molecular properties, like molar refraction, dipole moment, critical temperature, molar magnetic susceptibility, and cohesive energy. The set of descriptors is augmented with indicator variables for phenols, amines, nitriles, and aliphatic ring structures. Obviously, any QPPR model requires the experimental determination of the properties used as independent variables, and if such experimental values are missing, then the model is useless. Also, QPPR models cannot be used for property prediction of compounds not yet synthesized. The large number of experimental properties used as independent variables in the liquid viscosity QPPR models proposed by Suzuki makes their application difficult for a significant set of organic compounds whose properties, which are needed for the models, are not determined.

Usually, the liquid viscosity of organic compounds is computed with the corresponding states, hard sphere, square well, and modified Chapman−Enskog contribution methods.[1]

Molecular group contribution methods are widely employed to estimate the liquid viscosity.[4,5] The difficulty of this approach is represented by the definition of a consistent set of groups and by the necessity to compute the contribution of each group from a statistically significant number of molecules where the respective group is present. These methods are limited to molecules containing only the groups presented in the calibration set of molecules.

In the present investigation, we utilize a Hansch-type QSPR model that uses theoretical descriptors of the chemical structure. All QSPR and QSAR models are based on the assumption that the properties of a substance, like the physicochemical behavior, reactivity, or biological activity, are ultimately determined by its molecular structure. To correlate the molecular structure and properties, the chemical compounds must be adequately characterized with structural descriptors. The Hansch model[6,7] is a widely and effectively used QSAR approach that computes the biological activity with a multilinear regression (MLR) equation composed of terms of various physicochemical parameters assigned to the structures of the compounds. The technique is statistically based and intended at modeling molecular properties with structural descriptors. Using the least-squares method and statistical examinations, the regression coefficients of the MLR model are determined, specifying structural factors contributing to variations in the investigated property.

* To whom correspondence should be addressed.
[†] University "Politehnica" of Bucharest.
[‡] Romanian Academy.
[§] University of Nice−Sophia Antipolis.
[⊥] E-mail: o_ivanciuc@chim.upb.ro.

10.1021/ci980117v CCC: $18.00
Published on Web 03/31/1999

Originally, the Hansch model used steric, electronic, and hydrophobic descriptors, but nowadays, various constitutional, molecular graph (also named topological),[8] geometrical, electrostatic, charged surface area,[9] and quantum[10] descriptors are employed as independent variables in structure−property models.

CODESSA (comprehensive descriptors for structural and statistical analysis) represents a new QSPR approach that computes more than 500 structural parameters using the constitutional and quantum descriptions of the chemical compounds.[10−13] The computed structural descriptors are used to develop MLR models of the investigated property. CODESSA was successfully employed in QSPR studies concerning the prediction of gas chromatographic retention indices,[14] solubility of gases and vapors in water,[15,16] refractive indices,[17,18] boiling points,[19−21] melting points,[22] polymer glass transition temperatures,[23] solvent polarity scale,[24] and critical micelle concentration.[25] The main advantage of CODESSA over other statistical packages used in developing QSPR models is the easy generation of a large number of theoretical descriptors coding in various ways the chemical structure in a numerical form.

The purpose of this study is to develop, for the first time, a QSPR model for liquid viscosity prediction for a large variety of organic compounds using theoretic descriptors.

METHOD

Data. The 369 organic compounds used in developing the QSPR models, their experimental viscosity at 20 °C and literature sources, the calibration, prediction, and cross-validation residuals are presented in Table 1. The set of compounds is structurally very diverse and includes a large number of classes of organic compounds: alkanes, cycloalkanes, alkenes, cycloalkenes, aromatic hydrocarbons, alcohols, phenols, polyhydroxy alcohols, ethers, aldehydes, ketones, acids, esters, anhydrides, amides, nitriles, halogenated hydrocarbons, and nitro and sulfur compounds.

Structural Descriptors. The structures were drawn with HyperChem[26] and exported in a file format suitable for AMPAC.[27] The geometry optimization was performed on a 133-MHz Pentium with the semiempirical quantum method AM1[28] using the AMPAC 5.0 program. The HyperChem structure files and the AMPAC output files were used by the CODESSA program[29] to compute 579 descriptors. CODESSA computes five classes of structural descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.), topological (Wiener index, Randić connectivity indices, Kier shape indices, etc.), geometrical (moments of inertia, molecular volume, molecular surface area, etc.), electrostatic (when atomic charges are computed on the basis of atomic electronegativity: minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.), and quantum (minimum and maximum partial charges, Fukui reactivity indices, dipole moment, HOMO, LUMO, etc.).

Multilinear Regression Model. From the whole set of descriptors generated with CODESSA, we have discarded descriptors with a constant value for all molecules in the data set. Descriptors for which values were not available for every molecule were assigned a zero value for the missing position. CODESSA was used to develop the multilinear regression QSAR models by a heuristic method that considers the following steps:

(a) All quasi-orthogonal pairs of structural descriptors were selected from the initial set. Two descriptors are considered as being orthogonal if the intercorrelation coefficient (r_{ij}) is lower than 0.1.

(b) CODESSA uses the pairs of orthogonal descriptors to compute the biparametric regression equations. The most significant 10 pairs of molecular descriptors were used in the third step.

(c) To a MLR model built on n descriptors was added a new descriptor to generate a model with $n + 1$ descriptors if the new descriptor is not significantly correlated with the previous n descriptors (intercorrelation coefficient lower than 0.8).

Step c is repeated until MLR models with a certain maximum number of descriptors were obtained.

Model Validation. From the studies published over the last few years, the researchers become familiar with the importance of comparing the correlation (regression) models for prediction and not only for calibration (goodness of fit). The main goal in QSPR studies is to obtain a model with the highest predictive ability. Several techniques were proposed for comparing the quality of different models and for comparing the same model at different dimensionalities (number of structural descriptors).[30] An important problem in QSPR studies is the identification of structural descriptors that are to be used to model a given property. We have to point out here that correlation analysis develops models by suggesting statistical (and not causal) relationships between structural descriptors and a physical, chemical, or biological property. Correlations can be observed not only because a causal relationship exists between a set of descriptors and a property but also because of statistical bias resulting from errors in determining the structural descriptors, experimental errors in measuring the property, or even because of chance alone. When a correlation appears due to errors or chance factors, the resulting QSPR model has no scientific content and can lead to meaningless conclusions. As was shown by Topliss and co-workers,[31,32] such a situation can easily occur if large numbers of structural descriptors are screened for potential inclusion into the final correlation equation. Topliss and co-workers have studied the phenomenon of chance correlation by simulating correlations between randomly generated dependent and independent variables. They offered some specifications that can be used to estimate the occurrence of chance correlations at a given correlation coefficient for a specific combination of descriptors screened and experimental data. The problem of causality in selecting group contribution descriptors was investigated by Klopman.[33] Apart from the risk of chance correlations, the identification of the best QSPR equations for hundreds of experimental data and thousands of structural descriptors requires computational resources that are not presently available. A large variety of methods were proposed for descriptor selection,[34,35] such as genetic[36,37] and evolutionary algorithms.[38−40] Special algorithms for descriptor selection were developed for neural networks.[41−44]

Model validation techniques are used to distinguish between true and random correlations and to estimate the predictive power of the model.[30] The simplest model validation can be done by retaining the values of the

Table 1. Chemical Compounds, Experimental Viscosity, Calibration, and Prediction Residuals

no.	name	class[a]	$10^{-3}\,\eta$, Pa·s [$t = 20$ °C]	$\eta_{exp} - \eta_{calc}$	$\eta_{exp} - \eta_{cv}$	ref[b]	no.	name	class[a]	$10^{-3}\,\eta$, Pa·s [$t = 20$ °C]	$\eta_{exp} - \eta_{calc}$	$\eta_{exp} - \eta_{cv}$	ref[b]
1	decane	1	0.928	−0.026	−0.034	2	77	1-hexene	3	0.26	−0.035	−0.032	2
2	undecane	1	1.17	−0.026	−0.043	1	78	*cis*-3-hexene	3	0.273	−0.025	−0.022	5
3	dodecane	1	1.508	0.007	−0.023	2	79	*trans*-3-hexene	3	0.273	−0.025	−0.031	5
4	tridecane	1	1.883	0.002	0.023	2	80	1-heptene	3	0.35	−0.018	−0.015	2
5	tetradecane	1	2.18	−0.178	−0.253	1	81	1-octene	3	0.470	0.01	0.015	2
6	pentadecane	1	2.863	−0.09	−0.091	4	82	*trans*-3-octene	3	0.479	0.016	0.014	5
7	hexadecane	1	3.34	−0.359	−0.374	1	83	1-nonene	3	0.620	0.046	0.054	2
8	pentane	1	0.240	−0.074	−0.068	1	84	cyclopentene	4	0.346	0.056	0.061	3
9	2-methylbutane	1	0.225	−0.088	−0.084	3	85	cyclohexene	4	0.650	0.288	0.282	2
10	hexane	1	0.326	−0.056	−0.067	1	86	tetralin	5	2.202	0.821	0.804	2
11	2-methylpentane	1	0.310	−0.057	−0.057	2	87	butylbenzene	5	1.035	0.052	0.046	2
12	3-methylpentane	1	0.300	−0.142	−0.137	5	88	*sec*-butylbenzene	5	1.045	−0.108	−0.118	5
13	2,2-dimethylbutane	1	0.365	−0.018	−0.028	5	89	1-isopropyl-4-methylbenzene	5	3.402	2.282	2.302	2
14	2,3-dimethylbutane	1	0.376	−0.057	−0.068	5							
15	heptane	1	0.418	−0.064	−0.076	5	90	1-methylnaphthalene	5	3.629	1.726	1.744	5
16	2-methylhexane	1	0.378	−0.088	−0.081	2	91	*n*-pentylbenzene	5	1.33	0.103	0.088	4
17	3-methylhexane	1	0.372	−0.152	−0.146	2	92	*n*-hexylbenzene	5	1.67	0.137	0.159	4
18	2,4-dimethylpentane	1	0.361	−0.07	−0.062	2	93	*n*-heptylbenzene	5	2.07	0.156	0.164	4
19	2,2,3-trimethylbutane	1	0.579	0.02	0.01	2	94	*n*-octylbenzene	5	2.56	0.169	0.198	4
20	2,2-dimethylpentane	1	0.380	−0.049	−0.044	5	95	*n*-nonylbenzene	5	3.13	0.144	0.19	4
21	2,3-dimethylpentane	1	0.406	−0.18	−0.176	2	96	*n*-decylbenzene	5	3.79	0.061	−0.091	4
22	octane	1	0.542	−0.063	−0.076	1	97	*n*-undecylbenzene	5	4.59	−0.067	−0.284	4
23	2,2,3-trimethylpentane	1	0.598	−0.135	−0.135	2	98	*n*-dodecylbenzene	5	5.45	−0.365	−0.671	4
24	2,2,4-trimethylpentane	1	0.502	0.005	0.009	2	99	*n*-tridecylbenzene	5	6.48	−0.782	−0.959	4
25	2-methylheptane	1	0.521	−0.064	−0.062	5	100	*n*-tetradecylbenzene	5	7.66	−1.41	−1.643	4
26	3,3-dimethylhexane	1	0.500	−0.196	−0.196	5	101	*n*-pentadecylbenzene	5	8.99	−2.336	−3.137	4
27	3-ethylhexane	1	0.454	−0.236	−0.235	5	102	benzene	5	0.652	0.272	0.266	1
28	nonane	1	0.711	−0.047	−0.049	1	103	toulene	5	0.590	0.097	0.089	1
29	heptadecane	1	3.740	−0.89	−0.912	4	104	ethylbenzene	5	0.678	0.019	0.022	3
30	*trans*-decalin	2	2.128	0.439	0.417	2	105	*o*-xylene	5	0.810	0.057	0.07	1
31	*cis*-decalin	2	3.381	1.677	1.636	2	106	*m*-xylene	5	0.620	−0.046	−0.054	1
32	*n*-pentylcyclopentane	2	1.152	−0.114	−0.105	4	107	*p*-xylene	5	0.648	−0.014	−0.019	1
33	butylcyclohexane	2	1.31	0.056	0.057	4	108	styrene	5	0.751	0.044	0.054	2
34	*n*-hexylcyclopentane	2	1.880	0.297	0.304	4	109	1,3,5-trimethylbenzene	5	1.154	0.342	0.353	2
35	pentylcyclohexane	2	1.718	0.154	0.16	4	110	1,2,4-trimethylbenzene	5	1.116	0.178	0.156	5
36	bicyclohexyl	2	3.75	1.176	1.132	7	111	propylbenzene	5	0.855	0.074	0.07	4
37	*n*-heptylcyclopentane	2	2.360	0.377	0.362	4	112	isopropylbenzene	5	0.791	−0.044	−0.03	2
38	hexylcyclohexane	2	2.21	0.254	0.27	4	113	ethanol	6	1.200	−0.082	−0.02	1
39	*n*-octylcyclopentane	2	2.910	0.449	0.376	4	114	2-propyn-1-ol	6	1.68	0.308	0.364	3
40	heptylcyclohexane	2	2.80	0.35	0.363	4	115	allyl alcohol	6	1.363	−0.047	−0.035	3
41	*n*-nonylcyclopentane	2	3.550	0.476	0.523	4	116	1-propanol	6	2.234	0.458	0.426	4
42	octylcyclohexane	2	3.50	0.425	0.421	4	117	2-propanol	6	2.432	1.284	1.313	4
43	decylcyclopentane	2	3.55	−0.29	−0.27	4	118	1-butanol	6	2.948	0.721	0.832	1
44	nonylcyclohexane	2	4.31	0.449	0.523	4	119	2-butanol	6	3.907	1.952	2.016	2
45	undecylcyclopentane	2	4.28	−0.52	−0.751	4	120	2-methyl-1-propanol	6	4.034	1.816	1.818	4
46	decylcyclohexane	2	5.24	0.394	0.418	4	121	1-pentanol	6	4.00	1.262	1.14	4
47	dodecylcyclopentane	2	5.10	−0.923	−0.898	4	122	2-methyl-2-butanol	6	4.634	2.823	2.905	5
48	undecylcyclohexane	2	6.32	0.253	0.302	4	123	3-methyl-1-butanol	6	5.110	2.387	2.473	4
49	tridecylcyclopentane	2	6.04	−1.502	−1.417	4	124	2-methyl-1-butanol	6	5.50	2.319	2.332	3
50	dodecylcyclohexane	2	7.52	−0.075	−0.532	4	125	1-hexanol	6	5.040	1.5	1.676	5
51	tetradecylcyclopentane	2	7.11	−2.333	−2.021	4	126	1-heptanol	6	7.014	2.673	2.701	2
52	tridecylcyclohexane	2	8.89	−0.611	−1.24	4	127	2-heptanol	6	6.53	2.967	2.996	7
53	cyclopentane	2	0.439	0.047	0.049	2	128	benzyl alcohol	6	5.58	0.799	0.871	2
54	methylcyclopentane	2	0.507	−0.006	−0.001	2	129	methanol	6	0.597	−0.531	−0.539	1
55	cyclohexane	2	0.980	0.491	0.496	2	130	diisopentyl ether	8	1.012	−0.66	−0.7	2
56	methylcyclohexane	2	0.734	0.096	0.111	2	131	benzyl ether	8	5.33	−1.703	−2.135	6
57	ethylcyclohexane	2	0.843	−0.003	0.016	2	132	propylene oxide	8	0.327	−0.026	−0.032	2
58	*n*-propylcyclopentane	2	0.680	−0.126	−0.131	4	133	dimethoxymethane	8	0.325	−0.162	−0.161	2
59	*n*-butylcyclopentane	2	0.887	−0.126	−0.106	4	134	diethyl ether	8	0.245	−0.235	−0.234	3
60	*cis*,*cis*-1,3,5-trimethylcyclohexane	2	0.632	−0.388	−0.407	3	135	tetrahydrofuran	8	0.55	−0.025	−0.049	2
							136	1,2-epoxybutane	8	0.41	−0.049	−0.065	2
61	*trans*-1,3,5-trimethylcyclohexane	2	0.714	−0.301	−0.32	3	137	ethyl vinyl ether	8	0.2	−0.169	−0.18	3
							138	1,4-dioxane	8	1.20	0.445	0.441	3
62	propylcyclohexane	2	1.003	0.003	0.021	4	139	tetrahydropyran	8	0.826	0.124	0.131	3
63	1-decene	3	0.805	0.088	0.101	3	140	ethyl propyl ether	8	0.323	−0.278	−0.303	2
64	*n*-undecene	3	1.03	0.134	0.128	4	141	butyl vinyl ether	8	0.5	−0.076	−0.078	3
65	*n*-dodecene	3	1.30	0.182	0.21	4	142	butyl ethyl ether	8	0.421	−0.332	−0.331	2
66	*n*-tridecene	3	1.63	0.233	0.251	4	143	diisopropyl ether	8	0.341	−0.31	−0.304	3
67	*n*-tetradecene	3	2.00	0.256	0.269	4	144	1,2-diethoxyethane	8	0.65	−0.349	−0.366	2
68	*n*-pentadecene	3	2.46	0.281	0.201	4	145	bis(2-methoxyethyl) ether	8	1.99	0.717	0.788	3
69	*n*-hexadecene	3	2.99	0.27	0.333	4	146	methoxybenzene	8	1.32	0.247	0.231	1
70	*n*-heptadecene	3	3.60	0.203	0.225	4	147	ethoxybenzene	8	1.239	−0.037	−0.018	5
71	*n*-octadecene	3	4.31	0.066	0.19	4	148	dibutyl ether	8	0.691	−0.491	−0.514	5
72	1-pentene	3	0.197	−0.039	−0.037	5	149	acetaldehyde	9	0.244	−0.217	−0.219	3
73	2-methyl-1-butene	3	0.224	−0.025	−0.03	5	150	2-propenal	9	0.344	−0.362	−0.343	5
74	2-pentene (*cis* + *trans*)	3	0.214	−0.024	−0.018	7	151	propionaldehyde	9	0.404	−0.165	−0.148	3
75	2-methyl-1,3-butadiene	3	0.223	−0.043	−0.049	2	152	butyraldehyde	9	0.455	−0.267	−0.268	2
76	1,5-hexadiene	3	0.275	−0.07	−0.069	3	153	pentanal	9	0.513	−0.377	−0.378	5

Table 1. (Continued)

no.	name	class[a]	$10^{-3}\,\eta$, Pa·s [$t = 20$ °C]	$\eta_{exp} - \eta_{calc}$	$\eta_{exp} - \eta_{cv}$	ref[b]	no.	name	class[a]	$10^{-3}\,\eta$, Pa·s [$t = 20$ °C]	$\eta_{exp} - \eta_{calc}$	$\eta_{exp} - \eta_{cv}$	ref[b]
154	benzaldehyde	9	1.450	−1.082	−1.217	3	233	*N*,*N*-diethylaniline	15	2.18	−0.315	−0.387	1
155	2,6,8-trimethyl 4-nonanone	10	1.9	−0.758	−0.772	3	234	tributylamine	15	1.546	−1.685	−1.635	5
156	acetone	10	0.318	−0.01	−0.014	2	235	1,2-ethanediamine	15	1.54	0.098	0.145	2
157	2-butanone	10	0.428	−0.086	−0.084	2	236	butylamine	15	0.681	−0.135	−0.109	2
158	2-pentanone	10	0.506	−0.062	−0.062	4	237	diethylamine	15	0.355	−0.322	−0.334	5
159	3-pentanone	10	0.478	−0.247	−0.253	3	238	piperidine	15	1.487	0.270	0.314	5
160	cyclopentanone	10	1.168	0.36	0.367	5	239	pentylamine	15	1.018	−0.022	−0.004	3
161	2,4-pentanedione	10	0.6	−0.915	−0.944	3	240	cyclohexylamine	15	1.662	−0.046	−0.106	2
162	cyclohexanone	10	2.224	1.255	1.250	5	241	dipropylamine	15	0.534	−0.545	−0.530	2
163	4-methyl-2-pentanone	10	0.585	−0.090	−0.092	2	242	triethylamine	15	0.360	−0.645	−0.666	5
164	4-heptanone	10	0.736	−0.223	−0.220	2	243	aniline	15	4.40	2.676	2.727	1
165	acetophenone	10	1.810	−0.233	−0.289	3	244	*m*-toluidine	15	3.81	1.721	1.784	1
166	2,6-dimethyl-4-heptanone	10	1.03	−0.316	−0.313	3	245	methylphenylamine	15	2.360	0.231	0.129	5
167	acetic acid	11	1.232	−0.016	0.037	2	246	*o*-toluidine	15	4.39	2.290	2.276	1
168	acrylic acid	11	1.3	−0.823	−0.956	3	247	*N*,*N*-dimethylaniline	15	1.41	−0.411	−0.347	1
169	propionic acid	11	1.102	−1.027	−1.164	1	248	dibutylamine	15	0.95	−0.749	−0.752	3
170	methacrylic acid	11	1.32	−1.730	−1.819	2	249	nitroethane	16	0.677	0.090	0.135	2
171	butyric acid	11	1.540	−0.714	−0.831	1	250	1-nitropropane	16	0.844	0.167	0.163	7
172	isobutyric acid	11	1.311	−1.477	−1.388	5	251	2-nitropropane	16	0.770	−0.031	−0.029	7
173	*n*-valeric acid	11	2.24	−0.857	−0.951	4	252	nitrobenzene	16	2.03	0.576	0.539	1
174	2-ethylbutyric acid	11	3.3	−1.612	−1.937	3	253	2-nitrotoluene	16	2.37	0.399	0.357	1
175	*n*-caproic acid	11	3.20	−0.558	−0.381	4	254	3-nitrotoluene	16	2.33	0.449	0.404	1
176	*n*-heptylic acid	11	4.36	−0.610	−0.828	4	255	nitromethane	16	0.648	0.272	0.300	5
177	2-ethylhexanoic acid	11	7.7	0.421	0.795	3	256	tetranitromethane	16	1.76	−2.733	−2.964	3
178	octanoic acid	11	5.828	−0.203	−0.266	3	257	*N*,*N*-dimethylformamide	17	0.9243	−0.465	−0.538	7
179	*n*-nonylic acid	11	8.32	0.572	0.148	4	258	*N*,*N*-dimethylacetamide	17	2.141	1.101	1.061	2
180	formic acid	11	1.804	0.610	0.535	6	259	formamide	17	3.764	2.053	2.106	7
181	2-ethylhexyl acetate	12	1.5	−0.318	−0.293	3	260	1,1,2-trichloro-1,2,2-trifluoroethane	18	0.711	−1.535	−1.459	5
182	dibutyl maleate	12	5.62	0.285	0.404	3							
183	dibutyl *o*-phthalate	12	19.91	−0.461	−2.324	2	261	tetrachloroethylene	18	0.885	−0.219	−0.220	5
184	methyl formate	12	0.348	0.072	0.076	4	262	1,1,2,2-tetrabromoethane	18	9.797	3.029	4.187	2
185	methyl acetate	12	0.381	0.017	0.021	1	263	*cis*-1,2-dichloroethylene	18	0.467	0.029	0.034	2
186	ethyl formate	12	0.402	0.062	0.064	1	264	1,1-dichloroethylene	18	0.358	0.005	0.010	3
187	vinyl acetate	12	0.421	−0.041	−0.037	2	265	*trans*-1,2-dichloroethylene	18	0.404	−0.033	−0.026	3
188	methyl acrylate	12	1.398	0.846	0.858	3	266	1,2,2-trichloroethane	18	1.106	0.369	0.361	3
189	ethyl acetate	12	0.455	0.034	0.026	1	267	1,1,2-trichloroethane	18	0.119	−0.618	−0.601	3
190	propyl formate	12	0.574	0.150	0.156	2	268	1,2-dibromoethane	18	1.721	0.516	0.617	2
191	methyl propionate	12	0.477	−0.040	−0.049	3	269	1,1-dichloroethane	18	0.358	−0.012	−0.012	3
192	isopropyl formate	12	0.512	0.102	0.103	3	270	bromoethane	18	0.397	−0.005	0.010	2
193	propyl acetate	12	0.585	0.055	0.055	2	271	iodoethane	18	0.592	−0.031	−0.026	1
194	2-methylpropyl formate	12	0.680	0.165	0.164	2	272	1,1,2-trichloroethylene	18	0.566	−0.096	−0.107	2
195	methyl butyrate	12	0.580	−0.025	−0.026	2	273	pentachloroethane	18	2.45	0.511	0.625	3
196	methyl isobutyrate	12	0.523	−0.157	−0.144	2	274	allyl chloride	18	0.322	−0.050	−0.047	5
197	ethyl propionate	12	0.537	−0.069	−0.076	4	275	1,2-dichloropropane	18	0.865	0.261	0.271	2
198	*n*-butyl formate	12	0.691	0.161	0.162	3	276	1-chloropropane	18	0.352	−0.022	−0.019	1
199	isobutyl formate	12	0.680	0.165	0.165	3	277	2-chloropropane	18	0.400	0.092	0.082	5
200	isopropyl acetate	12	0.559	0.068	0.073	3	278	1-bromobutane	18	0.633	−0.048	−0.040	2
201	methyl methacrylate	12	0.632	−0.099	−0.115	3	279	2-bromobutane	18	1.434	0.731	0.730	2
202	methyl acetoacetate	12	1.704	0.421	0.427	3	280	1-bromo-2-methylpropane	18	0.643	−0.055	−0.063	2
203	ethyl acetoacetate	12	1.419	−0.083	−0.098	3	281	1-chlorobutane	18	0.458	−0.002	−0.001	5
204	2-methylpropyl acetate	12	0.697	0.051	0.058	3	282	2-chloro-2-methylpropane	18	0.512	0.160	0.150	5
205	ethyl butyrate	12	0.672	−0.038	−0.043	3	283	2-chlorobutane	18	0.412	−0.069	−0.069	5
206	butyl acetate	12	0.732	0.070	0.072	1	284	1-chloro-2-methylpropane	18	0.462	−0.015	−0.006	2
207	propyl propionate	12	0.673	−0.090	−0.086	2	285	1-iodo-2-methylpropane	18	0.875	−0.206	−0.201	2
208	methyl pentanoate	12	0.713	−0.048	−0.057	3	286	1-chloropentane	18	0.580	0.006	0.010	2
209	methyl maleate	12	3.54	1.965	2.010	7	287	1,3-dichlorobenzene	18	1.086	−0.500	−0.537	5
210	diethyl malonate	12	2.15	0.324	0.391	2	288	1,2-dichlorobenzene	18	1.428	−0.356	−0.367	5
211	pentyl acetate	12	0.924	0.098	0.100	2	289	bromobenzene	18	1.132	−0.299	−0.324	5
212	ethyl pentanoate	12	0.847	−0.046	−0.057	3	290	chlorobenzene	18	0.799	−0.167	−0.181	1
213	isopentyl acetate	12	0.872	0.069	0.077	3	291	fluorobenzene	18	0.598	−0.203	−0.224	1
214	2-methylbutyl acetate	12	0.872	−0.057	−0.065	3	292	iodobenzene	18	1.670	−0.522	−0.612	5
215	propyl butyrate	12	0.831	−0.062	−0.063	3	293	benzyl chloride	18	1.400	0.083	0.054	2
216	propyl isobutyrate	12	0.831	−0.180	−0.189	3	294	2-fluorotoluene	18	0.680	−0.585	−0.528	2
217	diethyl maleate	12	3.57	1.409	1.462	3	295	3-fluorotoluene	18	0.608	−0.494	−0.478	2
218	propyl pentanoate	12	1.053	−0.071	−0.092	3	296	4-fluorotoluene	18	0.622	−0.458	−0.483	2
219	methyl benzoate	12	2.044	0.120	0.199	5	297	carbon tetrachloride	18	0.969	0.340	0.348	1
220	ethyl benzoate	12	2.24	−0.023	−0.059	1	298	dibromomethane	18	1.018	0.297	0.299	5
221	acetic anhydride	13	0.907	0.304	0.302	2	299	bromochloromethane	18	0.670	0.190	0.194	3
222	propionic anhydride	13	1.144	−0.121	−0.107	3	300	dichloromethane	18	0.426	0.101	0.104	5
223	butyric anhydride	13	1.615	0.015	−0.003	3	301	diiodomethane	18	2.80	1.085	1.344	8
224	acetonitrile	14	0.375	0.073	0.078	3	302	iodomethane	18	0.500	−0.047	−0.013	2
225	acrylonitrile	14	0.35	−0.071	−0.084	2	303	chloroform	18	0.580	0.131	0.151	1
226	propionitrile	14	0.425	0.010	0.020	5	304	dimethyl sulfide	19	0.289	−0.032	−0.016	2
227	*trans*-3-butenenitrile	14	0.514	0.036	0.040	5	305	ethanethiol	19	0.250	−0.050	−0.038	5
228	methacrylonitrile	14	0.392	−0.085	−0.093	2	306	trimethylene sulfide	19	0.638	0.159	0.184	2
229	butyronitrile	14	0.616	0.118	0.129	5	307	ethyl methyl sulfide	19	0.373	−0.064	−0.062	3
230	4-methylvaleronitrile	14	0.980	0.225	0.238	7	308	diethyl sulfide	19	0.446	−0.102	−0.075	2
231	4-methylpentanenitrile	14	0.980	0.272	0.261	3	309	1-butanethiol	19	0.501	−0.008	−0.007	2
232	α-tolunitrile	14	2.161	0.510	0.555	7	310	thiophene	19	0.662	0.126	0.131	2

Table 1. (Continued)

no.	name	class[a]	$10^{-3}\ \eta$, Pa·s [$t = 20$ °C]	$\eta_{exp} - \eta_{calc}$	$\eta_{exp} - \eta_{cv}$	ref[b]	no.	name	class[a]	$10^{-3}\ \eta$, Pa·s [$t = 20$ °C]	$\eta_{exp} - \eta_{calc}$	$\eta_{exp} - \eta_{cv}$	ref[b]
311	tetrahydrothiophene	19	1.042	0.426	0.430	7	340	1-decanol	6	14.384		5.604	5
312	thiophenol	19	1.239	0.333	0.311	2	341	cyclohexanol	6	68.0		64.153	1
313	α-pinene	20	1.40	−0.231	−0.256	7	342	*m*-cresol	6	18.42		14.476	2
314	*β*-pinene	20	1.70	0.089	0.141	7	343	1-octanol	6	8.925		3.336	3
315	2,2,2-trifluoroethanol	20	1.996	−1.348	−1.211	2	344	2-ethyl-1-hexanol	6	9.8		3.728	3
316	acetaldoxime	20	1.415	−0.689	−0.64	3	345	1-nonanol	6	14.3		7.467	2
317	2-methoxyethanol	20	1.72	−0.618	−0.617	3	346	ethylene glycol	7	19.9		10.822	1
318	2-ethoxyethanol	20	2.04	−0.833	−0.739	3	347	1,2-propanediol	7	45.66		36.42	2
319	furan	20	0.380	−0.103	−0.12	2	348	1,3-propanediol	7	56.0		45.242	2
320	pyrrole	20	1.352	0.396	0.351	2	349	glycerol	7	1490		1429.262	1
321	methyl cyanoacetate	20	2.793	1.705	1.737	7	350	1,3-butanediol	7	130.3		118.391	3
322	2,2′-dichloroethyl ether	20	2.41	1.315	1.366	2	351	diethylene glycol	7	35.7		17.927	7
323	morpholine	20	2.23	0.637	0.606	3	352	1,5-pentanediol	7	128.0		110.996	2
324	tetrahydro-2-furanmethanol	20	6.24	1.513	1.717	8	353	2-methyl-2,4-pentanediol	7	34.4		24.782	3
325	pyridine	20	0.974	0.422	0.424	1	354	triethylene glycol	7	49.0		12.51	2
326	ethyl cyanoacetate	20	2.63	1.360	1.338	7	355	2-ethyl-1,3-hexanediol	7	323		294.94	3
327	1,2-ethanediol diacetate	20	3.13	1.859	1.863	2	356	oleic acid	11	38.80		−17.439	3
328	dicarbonic acid diethyl ester	20	1.97	−0.026	−0.023	8	357	ethyl cinnamate	12	8.7		5.235	2
329	4-hydroxy-4-methyl-2-pentanone	20	2.9	−0.135	−0.233	3	358	dibutyl decanedioate	12	9.03		−8.578	3
							359	methyl oleate	12	4.88		−8.554	3
330	2,2-dimethyl-1,3-dioxolane-4-methanol	20	11	2.857	3.163	3	360	maleic anhydride	12	4.436		3.623	5
							361	*N*-methylpropionamide	17	6.06		4.569	2
331	2-(2-ethoxyethoxy)ethanol	20	3.85	−2.171	−2.106	7	362	2,2′-thiodiethanol	19	65.2		46.529	3
332	3-bromoaniline	20	6.81	2.659	2.668	1	363	ethanolamine	20	23.22		18.966	5
333	*o*-chloroaniline	20	2.916	−0.004	−0.123	7	364	trifluoroacetic acid	20	0.926		−3.835	2
334	2-methylpyridine	20	0.805	−0.133	−0.165	3	365	glycerol trinitrate	20	36.0		30.924	3
335	benzoyl bromide	20	1.956	−0.677	−0.789	2	366	bis(2-hydroxyethyl)ether	20	35.7		17.932	3
336	2,4,6-trimethylpyridine	20	1.498	−0.04	0	2	367	tetrahydropyran-2-methanol	20	11.0		4.871	3
337	4-*tert*-butylpyridine	20	1.495	−0.351	−0.303	3							
338	pentadecylcyclopentane	2	8.3		−3.521	4	368	salicyl aldehyde	20	2.90		−4.788	3
339	hexadecylcyclopentane	2	9.6		−5.19	4	369	salicylic acid	20	2.71		−23.22	6

[a] Classes of compounds: 1, alkanes; 2, cycloalkanes; 3, alkenes; 4, cycloalkenes; 5, aromatic hydrocarbons; 6, alcohols and phenols; 7, polyhydroxy alcohols; 8, ethers; 9, aldehydes; 10, ketones; 11, acids; 12, esters; 13, anhydrides; 14, nitriles; 15, amines; 16, nitro compounds; 17, amides; 18, halogenated hydrocarbons; 19, sulfur compounds; 20, other. [b] References: (1) Weast, R. C., Ed. *Handbook of Chemistry and Physics*, 69th ed.; CRC: Boca Raton, FL, 1988−1989. (2) Dean, J. A. *Handbook of Organic Chemistry*; McGraw-Hill: New York, 1987. (3) Dean, J. A. *Lange's Handbook of Organic Chemistry*, 14th ed.; McGraw-Hill: New York, 1992. (4) Pachaiyappan, V. S.; Ibrahim, S. H.; Kuloor, N. R. Simple Correlation for Determining Viscosity of Organic Liquids. *Chem. Eng.* **1967**, *74*, 193−196. (5) Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *The Properties of Gases and Liquids*, 4th ed.; McGraw-Hill: New York, 1987. (6) Weast, R. C., Ed. *Handbook of Chemistry and Physics*, 65th ed.; CRC: Boca Raton, FL, 1984−1985. (7) Riddick, J. A.; Bunger, W. A. *Organic Solvents Physical Properties and Methods of Purification*; Wiley-Interscience: New York, 1970. (8) The Merck Index, 12th ed.; CD-ROM, version 12:1, Chapman & Hall: London, 1996.

dependent variables (experimental property) assigned to the corresponding compounds and randomly generating the independent variables (structural descriptors).[41,42,45,46] If the fit obtained with the real set is consistently better than that with the random independent variables, the correlation obtained with the real data can confidently be said not to be due to chance factors.

Because the simulations using random independent variables give no idea of the effects of correlation between the independent variables and inhomogeneous distribution of the values of parameters, the second method uses a data set obtained from the real one by randomly assigning a real property value to a compound not necessarily possessing this property and by retaining the values of the independent variables.[45,46] This procedure is repeated a number of times, and if the fit obtained with the real data set is constantly better than the one obtained with randomly assigned property data, one can be confident that a valid relationship has indeed been established between the structural descriptors and the experimental data.

The third method, cross-validation, is the most commonly used technique for evaluating the predictive ability of a model for a given data set. In the first step, the data set S that contains N elements is divided into n groups G_i. In the second step, the model is calibrated with a data set obtained from S by eliminating the first cross-validation group of data G_1. In the third step, the coefficients of the calibrated model are used to predict the investigated property for all compounds in the set G_1. The second and third steps are repeated n times, until each compound is selected once in a cross-validation group and $n - 1$ times in calibration groups. By comparing the experimental and predicted value of the modeled property, one obtains a measure of the predictive power of the model for a given set of structural descriptors and for a given data set. One can increase the number of groups n until it equals the number of data points N, giving the leave-one-out cross-validation procedure. However, this technique is not recommended for large data sets, when the computational effort becomes too high.

Although the QSPR equations developed with CODESSA are obtained by selection of descriptors from a large pool, several descriptor selection techniques were used in order to minimize the possibility of chance correlations. In the first step, from the initial pool of descriptors, CODESSA eliminates all descriptors with F-test values of less than 1, t-test values of less than 0.1, or correlation coefficients of less than 0.1, thus greatly reducing the dimensionality of the problem—that of finding a QSPR equation with a good predictive power. Then, as described in the previous section, a heuristic algorithm selects only quasi-orthogonal sets of descriptors that are tested for correlation with the viscosity data. This selection algorithm ensures that the probability

Table 2. Notation of the Descriptors Involved in the QSPR Models with Five Theoretical Parameters

no.	type[a]	notation	descriptors
1	E	**A**	PPSA-3, atomic charge weighted partial positive surface area (electrostatic)
2	E	**B**	topographic electronic index (all bonds) (electrostatic)
3	E	**C**	HDCA-2, hydrogen bonding donor charged surface area (electrostatic)
4	E	**D**	HDCA-1, hydrogen bonding donor charged surface area (electrostatic)
5	Q	**E**	HDSA, hydrogen-donors surface area (quantum)
6	E	**F**	topographic electronic index (all pairs) (electrostatic)
7	Q	**G**	HDCA, hydrogen-donor charged surface area (quantum)
8	E	**H**	HDCA-2/SQRT(TMSA), fractional HDCA-2 (electrostatic)
9	Q	**I**	HDCA-2, hydrogen bonding donor charged surface area (quantum)
10	Q	**J**	HDCA-1, hydrogen bonding donor charged surface area (quantum)
11	C	**K**	gravitation index (all bonds)
12	C	**L**	gravitation index (all pairs)
13	C	**M**	molecular weight
14	E	**N**	HDCA-2/TMSA, fractional HDCA-2 (electrostatic)
15	T	**O**	Randić index (order 3)
16	T	**P**	Randić index (order1)
17	E	**Q**	HACA-2, hydrogen bonding acceptor charged surface area (electrostatic)
18	C	**R**	no. of rings
19	C	**S**	no. of O atoms
20	Q	**T**	max electrophilic reactivity index for a C atom
21	C	**U**	rel no. of rings
22	T	**V**	structural information content (order 0)
23	Q	**W**	FNSA-3, fractional charged negative surface area (quantum)
24	Q	**X**	FHASA, fractional surface area of hydrogen acceptor atoms (quantum)
25	Q	**Y**	max atomic orbital electronic population
26	C	**Z**	rel no. of double bonds
27	Q	**Aa**	max net atomic charge for a O atom
28	Q	**Ab**	av electrophilic reactivity index for a C atom

[a] Descriptor type: C, constitutional; T, topological; G, geometrical; E, electrostatic; Q, quantum.

of obtaining a chance correlation is low and maintains a reasonable searching time. We have to mention that the heuristic nature of the algorithm offers only near-optimal solutions, because the exhaustive search of the best QSPR equation for 369 experimental data and 579 descriptors requires very large computational resources.

In CODESSA, the model validation is performed with the leave-one-out cross-validation procedure. However, for MLR equations, this technique induces a too small modification of the initial data set, and for the present study, we have estimated the stability and predictive ability of the MLR model with the leave-20%-out (L20%O) cross-validation method.[47−49] The L20%O cross-validation was applied by establishing a prediction set consisting of 20% of the patterns randomly selected from the entire set of experimental data and retaining the remaining 80% of the patterns in a calibration set. The MLR model is developed with the calibration set, and in a second phase, its coefficients are used to compute the viscosity of the molecules from the prediction set. This procedure is repeated five times, until all molecules are selected in a prediction set once and only once. A linear regression between experimental ln η_{exp} and predicted ln η_{pr} offers the statistical indices used to compare the stability and predictive capacity of the different MLR models. Because CODESSA does not offer this model validation option, the L20%O cross-validation was performed by exporting the data from CODESSA and doing the computations with the programs used in other QSPR investigations.[47−49]

RESULTS AND DISCUSSION

Using the HyperChem structural files and AMPAC AM1 output files for 369 compounds, CODESSA generated a set of 579 descriptors. Table 2 presents the notation of the 28 descriptors involved in the correlations reported in this paper. The definitions of the descriptors are presented in detail in the CODESSA Reference and User's Manuals.[29]

Table 3. One-Parameter Regression for Computing the Viscosity

no.	descriptor	*r*	*F*
1	**A**	0.665	17
2	**B**	0.634	16
3	**C**	0.625	15
4	**D**	0.624	15
5	**E**	0.615	15
6	**F**	0.604	16
7	**G**	0.602	14
8	**H**	0.593	14
9	**I**	0.587	14
10	**J**	0.584	14

The viscosity values (η) of the compounds involved in the present study span a large interval, and as in other QSPR viscosity studies, we have used ln η as the experimental value to restrain the range of property values. The theoretical parameters computed with CODESSA were used in monoparametric correlations relating ln η to the structural descriptors, and those offering the best results were selected for further use. The number of descriptors was reduced to 392 by eliminating descriptors with *F*-test values of less than 1, *t*-test values of less than 0.1, or correlation coefficients of less than 0.1. Table 3 presents the best 10 monoparametric correlations, giving the descriptor involved, the correlation coefficient (*r*) and the *F*-test value. Six from the best 10 descriptors are electrostatics and the remaining four are quantum.

The biparametric correlations were computed by using all pairs of descriptors with an intercorrelation coefficient of lower than 0.8. The best 10 biparametric correlations are

Table 4. Biparameter Regression for Computing the Viscosity

no.	descriptors	*r*	*F*
1	**C, K**	0.892	711
2	**C, L**	0.888	685
3	**C, M**	0.880	627
4	**I, K**	0.875	599
5	**I, L**	0.872	582
6	**N, K**	0.868	561
7	**I, M**	0.867	553
8	**C, O**	0.865	545
9	**Q, K**	0.861	524
10	**C, P**	0.860	522

presented in Table 4. The introduction of a second parameter significantly improves the statistical indices of the QSPR model, with the correlation coefficient ranging between 0.86 and 0.89 and the *F* test between 522 and 711. The top three correlations involve the hydrogen-acceptor index **C** and a constitutional descriptor: bond gravitation index **K**, pair gravitation index **L**, and molecular weight **M**, respectively. The three constitutional descriptors **K**, **L**, and **M** are related, because **K** and **L** are computed on the basis of the atomic weights and geometric distances separating the atoms. Equations 4, 5, and 7 from Table 4 involve the hydrogen-acceptor index **I** (with the same definition as index **C** but computed with quantum charges) and one of the constitutional descriptors **K**, **L**, and **M**, respectively. The hydrogen-acceptor index **C** and a Randić connectivity index, either **O** or **P**, are used in eqs 8 and 10. Equation 9 involves another electrostatic index, **Q**, and the gravitation index **K**. As a general trend, the biparametric correlations involve an electrostatic index and another index describing the molecular weight, weight distribution, molecular shape, or connectivity.

Using the set of biparametric correlations obtained in the previous step, a heuristic method was applied to generate correlations with up to five parameters. The best 10 such correlations are presented in Table 5. The introduction of five parameters significantly improves the statistical indices of the QSPR model, with the correlation coefficient being between 0.915 and 0.919 and the *F* test between 374 and 394. All best 10 correlations involve the hydrogen-acceptor index **C** and molecular weight **M**. The statistical indices (correlation coefficient *r*, standard deviation *s*, Fisher's test *F*) of the MLR equations relating ln η to the structural descriptors are presented in Table 5, columns 12−14. The *r* values of the 10 QSPR models from Table 5 indicate a good correlation between ln η and the structural descriptors, but for practical applications, it is important to have a good prediction of the η values themselves. Columns 15 and 16 from Table 5 give the *r* and *s* statistical indices of the equation

$$\eta_{exp} = a + b\eta_{MLRcal} \quad (1)$$

where η_{MLRcal} represents the viscosity value computed with one of the QSPR models reported in Table 5. The *r* values from column 15 indicate that the estimation of the viscosity is not as good as was concluded on the basis of the statistical indices of the ln η QSPR. By comparing the statistical indices of the 10 QSPR equations reported in Table 5, it is clear that the statistical differences between them are small, and it is not easy to select univocally the best equation from this set. Two descriptors are common to all 10 equations, namely, **C**, the hydrogen-bonding donor charged surface HDCA-2 computed with atomic charges derived from atomic electronegativities, and **M**, the molecular weight that represents a molecular size descriptor. The most important parameter, HDCA-2, is a descriptor directly related to the hydrogen bonding between the molecules in condensed media and is defined by

$$\text{HDCA-2} = \sum \frac{Q_d SA_d^{1/2}}{MSA^{1/2}} \quad (2)$$

where Q_d is the partial charge on the hydrogen bonding donor atoms, SA_d denotes the exposed surface area of this atom, and MSA is the total molecular surface area calculated from the van der Waals radii of the atoms. The summation in the formula of HDCA-2 goes over all possible hydrogen-bonding donor atoms in a molecule.

The presence of HDCA-2 and the molecular weight descriptors in all equations reported in Table 5 demonstrates that the viscosity of organic liquids is determined mainly by the propensity of forming hydrogen bonds and by the molecular size. Several descriptors have a high incidence in the set of 10 equations from Table 5: **R**, the number of rings, appears in six equations; **S**, the number of oxygen atoms, appears in six equations; **T**, the maximum electrophilic reactivity index for a carbon atom, appears in six equations; **U**, the relative number of rings, appears in three equations; **Aa**, the maximum net atomic charge for an oxygen atom, appears in two equations. The good statistical quality of the equations from Table 5 demonstrates that the descriptor selection algorithms from CODESSA are able to extract from a large set of descriptors a small group that influence the viscosity of organic compounds.

An inspection of the residuals (difference between η_{exp} and η_{MLRcal}) of the 10 QSPR models shows that in each of them there are a number of outliers (compounds with a residual larger than 3*s*). To select the best QSPR equation, after the elimination of the outliers and the L20%O cross-validation test, the statistical indices of eq 1 were determined and the equation with the best prediction statistical indices was selected. The best QSPR model selected for further investigation after the elimination of the outliers is derived from eq 8 in Table 5:

$$\ln \eta_{exp} = -2.814(\pm 0.144) + 3.387(\pm 0.139)\mathbf{C} + 8.858 \times 10^{-3}(\pm 0.647 \times 10^{-3})\mathbf{M} + 3.919 \times 10^{-1}(\pm 0.303 \times 10^{-1})\mathbf{O} - 8.486(\pm 1.435)\mathbf{T} + 6.684 \times 10^{-1}(\pm 0.803 \times 10^{-1})\mathbf{Y} \quad (3)$$

$$n = 337 \quad r = 0.920 \quad s = 0.371 \quad F = 367$$

In this equation, the standard error of estimation of each coefficient at the 95% confidence level is given within parentheses. The model contains the following five theoretical parameters: **C**, the hydrogen-bonding donor charged surface HDCA-2 computed with atomic charges derived from atomic electronegativities; **M**, the molecular weight; **O**, the Randić connectivity index of order 3, $^3\chi$; **T**, the maximum electrophilic reactivity index for a carbon atom; and **Y**, the maximum atomic orbital electronic population.

Table 5. Multiparameter Regressions for Computing the Viscosity[a]

no	a_0	a_1	p_1	a_2	p_2	a_3	p_3	a_4	p_4	a_5	p_5	$r(\ln \eta)$	$s(\ln \eta)$	$F(\ln \eta)$	$r(\eta)$	$s(\eta)$
1	−1.894	4.313	**C**	0.014	**M**	0.510	**R**	0.178	**S**	−10.045	**T**	0.919	0.516	394	0.817	7.920
2	−1.995	4.151	**C**	0.014	**M**	0.489	**R**	−1.473	**Aa**	−10.621	**T**	0.917	0.522	383	0.793	6.448
3	−1.934	4.323	**C**	0.015	**M**	8.037	**U**	−11.639	**T**	0.161	**S**	0.916	0.523	382	0.802	8.361
4	−2.161	4.502	**C**	0.012	**M**	7.231	**U**	−12.742	**T**	0.124	**V**	0.916	0.525	378	0.793	7.681
5	−2.039	4.159	**C**	0.016	**M**	8.010	**U**	−12.131	**T**	−1.405	**Aa**	0.916	0.526	377	0.775	6.959
6	−2.022	4.540	**C**	0.013	**M**	0.636	**R**	0.261	**S**	5.380	**W**	0.916	0.526	376	0.798	8.350
7	−2.277	5.026	**C**	0.014	**M**	0.619	**R**	0.172	**S**	−2.108	**X**	0.915	0.526	376	0.835	7.904
8	−2.765	4.147	**C**	0.009	**M**	0.366	**O**	−13.808	**T**	0.731	**Y**	0.915	0.526	376	0.818	6.906
9	−2.149	4.356	**C**	0.013	**M**	0.580	**R**	0.188	**S**	−2.091	**Z**	0.915	0.526	376	0.822	7.822
10	−2.051	4.437	**C**	0.014	**M**	0.567	**R**	0.151	**S**	−12.007	**Ab**	0.915	0.527	374	0.803	7.799

[a] p_1-p_5 represent structural descriptors; a_0-a_5 represent the coefficients of the MLR model; $r(\ln \eta)$, $s(\ln \eta)$, and $F(\ln \eta)$, are the statistical indices for ln η QSPR; $r(\eta)$ and $s(\eta)$ are the statistical indices for the corresponding equation (1).

The third descriptor in eq 3, the Randić connectivity index of order 3, collects the weighted contribution of clusters formed by four atoms and is a measure of molecular branching. The remaining two descriptors are derived from AM1 quantum computations, namely, the maximum electrophilic reactivity index for a carbon atom and the maximum atomic orbital electronic population. The electrophilic reactivity index for a given atomic species A is defined as

$$E_A = \frac{\sum_{i=1}^{n_A} C_{\mathrm{LUMO},i}{}^2}{\epsilon_{\mathrm{LUMO}} + 10} \tag{4}$$

where the summation is performed over all valence atomic orbitals i in atom A ($i = 1, ..., n_A$), $C_{\mathrm{LUMO},i}$ denote the ith AO coefficient on the lowest unoccupied molecular orbital (LUMO), and ϵ_{LUMO} is the energy of this orbital. The maximum atomic orbital electronic population is an index that describes the nucleophilicity of the molecule. The examination of the coefficients of the **T** and **Y** descriptors in eq 3 shows that the viscosity decreases with an increase of the electrophilicity of the carbon atom and increases with the increase of the molecular nucleophilicity.

If we transform the ln η_{MLRcal} values calculated with eq 3 into the viscosity values η_{MLRcal}, we obtain the following equation:

$$\eta_{\mathrm{exp}} = 0.087 + 0.904\eta_{\mathrm{MLRcal}} \tag{5}$$
$$n = 337 \quad r = 0.931 \qquad s = 0.746$$

This calibration equation gives a good estimation of the viscosity, with a fairly small standard deviation. The calibration residuals of the 337 compounds are presented in Table 1, column 5. Column 6 from the same table gives the residual viscosity of the 32 outliers predicted with eq 3. The set of outliers contains mainly esters and diols, some higher alcohols, glycerol, oleic acid, 2,2′-thiodiethanol, ethanolamine, bis(2-hydroxyethyl) ether, and salicylic acid. Other outliers have residuals at the limit of 3s, and therefore, it is difficult to say if they represent true outliers or if their experimental viscosities are in error. The chemical structure of the outliers indicates that the QSPR model is not able to predict the viscosity of highly polar organic compounds, containing two or more polar groups. New structural descriptors have to be developed for such compounds.

From a practical point of view, a QSPR model is as valuable as its predictions are. The L20%O cross-validation was applied to the model from eq 3, giving the following results:

$$\ln \eta_{\mathrm{exp}} = 0.003 + 0.992 \ln \eta_{\mathrm{MLRcv}} \tag{6}$$
$$n = 337 \qquad r_{\mathrm{cv}} = 0.917 \qquad s_{\mathrm{cv}} = 0.377$$
$$\eta_{\mathrm{exp}} = 0.053 + 0.929\eta_{\mathrm{MLRcv}} \tag{7}$$
$$n = 337 \qquad r_{\mathrm{cv}} = 0.922 \qquad s_{\mathrm{cv}} = 0.818$$

The cross-validation residuals of the 337 compounds are presented in Table 1, column 6. As is apparent from the cross-validation residuals and the statistical indices of eqs 6 and 7, the predictions offered by the QSPR model from eq 3 are good, with a low decrease of the correlation coefficient r_{cv} and a small increase of the standard deviation s_{cv}, compared with the corresponding indices from eqs 3 and 5. The cross-validation results show that although the QSPR equations developed with CODESSA are obtained by choosing descriptors from a large pool, the descriptor selection algorithms and the heuristic generation of the MLR models minimize the possibility of chance correlations and provide models with a good predictive power. QSPR models containing more than five parameters were investigated and showed no improvement over the results reported in the above equations.

CONCLUSION

A QSPR model for the estimation of the liquid viscosity of a large variety of organic compounds was established using the CODESSA program. The final model, developed with a calibration set containing 337 compounds, has good statistical indices; i.e., $s = 0.37$ and $r = 0.920$. Five theoretical parameters were included in the QSPR model: molecular weight; Randić connectivity index of order 3; hydrogen-donor charged surface area HDCA-2 (electrostatic); maximum electrophilic reactivity index for a C atom; maximum atomic orbital electronic population. The predictive ability of the MLR model was tested by the leave-20%-out cross-validation method, showing that the QSPR model is stable and can be used to obtain good predictions for compounds that were not used in the model calibration. The cross-validation statistical indices show a small decrease when compared with those obtained in the calibration phase; i.e., $s_{\mathrm{cv}} = 0.38$ and $r_{\mathrm{cv}} = 0.917$. Although the QSPR equations

developed with CODESSA are obtained by selection of descriptors from a large pool, we have used several descriptor selection techniques in order to minimize the possibility of chance correlations. The QSPR models developed with CODESSA allow accurate computation of the liquid viscosity of organic compounds using simple constitutional descriptors and quantum indices that can be computed with standard quantum chemistry packages. The application of the model is limited to compounds with one polar group, while for the computation of the viscosity for highly polar compounds, new structural parameters have to be developed.

ACKNOWLEDGMENT

O.I. thanks the Ministère de l′Education Nationale, de l′Enseignement Supérieur et de la Recherche of France, for a PAST grant. O.I. also acknowledges a software grant consisting of HyperChem Release 5 offered by Hypercube, Inc., and a software grant consisting of AMPAC 5.0 and CODESSA 2.13 offered by Semichem. We acknowledge the partial financial support of T.I. and O.I. by the Ministry of Research and Technology under Grant 381 TA10 and by the Ministry of Education under Grant 5001 TB10.

REFERENCES AND NOTES

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

(2) Suzuki, T.; Ohtaguchi, K.; Koide, K. Computer-Assisted Approach to Develop a New Prediction Method of Liquid Viscosity of Organic Compounds. *Comput. Chem. Eng.* **1996**, *20*, 161−173.

(3) 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.

(4) Joback, K. G.; Reid, R. C. Estimation of Pure-Component Properties from Group-Contributions. *Chem. Eng. Commun*. **1987**, *57*, 233−243.

(5) Škubla, P. Prediction of Viscosity of Organic Liquids. *Collect. Czech. Chem. Commun*. **1985**, *50*, 1907−1916.

(6) Hansch, C.; Fujita, T. $\rho-\sigma-\pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc*. **1964**, *86*, 1616−1626.

(7) Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, pi, Derived from Partition Coefficients. *J. Am. Chem. Soc*. **1964**, *86*, 5175−5180.

(8) Ivanciuc, O.; Balaban, A. T. Graph Theory in Chemistry. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, 1998; pp 1169−1190.

(9) 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.

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

(11) 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*, 17−23.

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

(13) Katritzky, A. R.; Lobanov, V. S.; Karelson, M.; Murugan, R.; Grendze, M. P.; Toomey, J. E., Jr. Comprehensive Descriptors for Structural and Statistical Analysis. 1. Correlations Between Structure and Physical Properties of Substituted Pyridines. *Rev. Roum. Chim*. **1996**, *41*, 851−867.

(14) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. 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.

(15) Katritzky, A. R.; Mu, L.; Karelson, M. A QSPR Study of the Solubility of Gases and Vapors in Water. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 1162−1168.

(16) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water−Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720−725.

(17) 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.

(18) Katritzky, A. R.; Sild, S.; Karelson, M. Correlation and Prediction of the Refractive Indices of Polymers by QSPR. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1171−1176.

(19) 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.

(20) 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.

(21) Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Quantitative Structure−Property Relationship Study of Normal Boiling Points for Halogen-/Oxygen-/Sulfur-Containing Organic Compounds Using the CODESSA Program. *Tetrahedron* **1998**, *54*, 9129−9142.

(22) Katritzky, A. R.; Maran, U.; Karelson, M.; Lobanov, V. S. Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 913−919.

(23) Katritzky, A. R.; Rachwal, P.; Law, K. W.; Karelson, M.; Lobanov, V. S. Prediction of Polymer Glass Transition Temperatures Using a General Quantitative Structure−Property Relationship Treatment. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 879−884.

(24) Katritzky, A. R.; Mu, L.; Karelson, M. QSPR Treatment of the Unified Nonspecific Solvent Polarity Scale. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 756−761.

(25) Huibers, P. D. T.; Lobanov, V. S.; Katritzky, A. R.; Shah, D. O.; Karelson, M. Prediction of Critical Micelle Concentration Using a Quantitative Structure−Property Relationship Approach. 1. Nonionic Surfactants. *Langmuir* **1996**, *12*, 1462−1470.

(26) HyperChem 4.5, Hypercube, Inc., 419 Phillip St., Waterloo, Ontario, Canada N2L 3X2; Telephone: (519) 725-4040. Fax: (519) 725-5193. Information hot-line: (800) 960-1871. E-mail information requests: info@hyper.com. E-mail URL: http://www.hyper.com.

(27) AMPAC 5.0, Semichem, 7128 Summit, Shawnee, KS 66216. E-mail: aholder@cctr.umkc.edu.

(28) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc*. **1985**, *107*, 3902−3909.

(29) CODESSA 2.13, Semichem, 7128 Summit, Shawnee, KS 66216. E-mail: aholder@cctr.umkc.edu.

(30) Wold, S.; Eriksson, L. Validation Tools. In *QSAR: Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; Methods and Principles in Medicinal Chemistry, Vol. 2, Verlag Chemie: Weinheim, Germany, 1995; pp 309−318.

(31) Topliss, J. G.; Costello, R. J. Chance Correlations in Structure−Activity Studies Using Multiple Regression Analysis. *J. Med. Chem*. **1972**, *15*, 1066−1078.

(32) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem*. **1979**, *22*, 1238−1244.

(33) Klopman, G.; Kalos, A. N. Causality in Structure−Activity Studies. *J. Comput. Chem*. **1985**, *6*, 492−506.

(34) Livingstone, D. J.; Rahr, E. CORCHOP—An Interactive Routine for the Dimension Reduction of Large QSAR Data Sets. *Quant. Struct.-Act. Relat*. **1989**, *8*, 103−108.

(35) McFarland, J. W.; Gans, D. J. On Identifying Likely Determinants of Biological Activity in High Dimensional QSAR Problems. *Quant. Struct.-Act. Relat.* **1994**, *13*, 11−17.

(36) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom*. **1992**, *6*, 267−281.

(37) Leardi, R. Application of a Genetic Algorithm to Feature Selection Under Full Validation Conditions and to Outlier Detection. *J. Chemom*. **1994**, *8*, 65−79.

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

(39) Kubinyi, H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.

(40) Kubinyi, H. Evolutionary Variable Selection in Regression and PLS Analyses. *J. Chemom*. **1996**, *10*, 119−133.

(41) Manallack, D. T.; Livingstone, D. J. Artificial Neural Networks: Application and Chance Effects for QSAR Data Analysis. *Med. Chem. Res*. **1992**, *2*, 181−190.

(42) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem*. **1993**, *36*, 1295−1297.
(43) Wikel, J. H.; Dow, E. R. The Use of Neural Networks for Variable Selection in QSAR. *Bioorg. Med. Chem. Lett*. **1993**, *3*, 645−651.
(44) Tetko, I. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci*. **1996**, *36*, 794−803.
(45) Ivanciuc, O. Artificial Neural Networks Applications. Part 4. Quantitative Structure−Activity Relationships for the Estimation of the Relative Toxicity of Phenols for *Tetrahymena*. *Rev. Roum. Chim*. **1998**, *43*, 255−260.
(46) Ivanciuc, O. Artificial Neural Networks Applications. Part 7. Estimation of Bioconcentration Factors in Fish Using Solvatochromic Parameters. *Rev. Roum. Chim*. **1998**, *43*, 347−354.
(47) Ivanciuc, O.; Rabine, J.-P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. ^{13}C NMR Chemical Shift Prediction of sp^2 Carbon Atoms in Acyclic Alkenes using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644−653.
(48) Ivanciuc, O.; Rabine, J.-P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. ^{13}C NMR Chemical Shift Prediction of the sp^3 Carbon Atoms in the α Position Relative to the Double Bond in Acyclic Alkenes. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 587−598.
(49) Ivanciuc, O.; Rabine, J.-P.; Cabrol-Bass, D. ^{13}C NMR Chemical Shift Sum Prediction for Alkanes using Neural Networks. *Comput. Chem.* **1997**, *21*, 437−443.

CI980117V