Correlation of liquid viscosity with molecular structure for organic compounds using different variable selection methods

Bono Lučić,[1*] Ivan Bašic,[2] Damir Nadramija,[2] Ante Miličević,[1] Nenad Trinajstić,[1] Takahiro Suzuki,[3] Ruslan Petrukhin,[4,5] Mati Karelson[4] and Alan R. Katritzky[5]

[1]*The Rugjer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia*
[2]*PLIVA, Pharmaceutical industry, Research Information Center, Prilaz Baruna Filipovića 25, HR-10000 Zagreb, Croatia*
[3]*Research Laboratory of Resources Utilization, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226, Japan*
[4]*Department of Chemistry, University of Tartu, 2 Jakobi Str., EE51014 Tartu, Estonia*
[5]*Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, P.O.Box 117200, Gainesville, Florida 32611-7200, U. S. A.*
E-mail: lucic@faust.irb.hr

Abstract

Improved models for predicting viscosities at 20 °C were generated using three different methods for descriptor selection. Data set of 361 diverse organic molecules and their experimental viscosities were used for developing the models. Molecular properties are encoded by 822 initial descriptors computed by the CODESSA program. CODESSA, GFA and CROMRsel methods are capable of selecting good and facile viscosity models having only five descriptors. These methods are automated procedures for generation of simple multiregression (MR) models. All three methods produce excellent linear models, but the models obtained by the CROMRsel method are somewhat better. In addition, using the CROMRsel suite of programs a very good nonlinear MR model having five descriptors (two linear and three cross-product descriptors, R^2 = 0.908, S = 0.175) was obtained. Nonlinear models generated in this study show that the classical MR based methods can be efficiently used to obtain simple and very good nonlinear MR models. The best five-descriptor models selected in this study usually contain one geometrical (gravitational index) and one topological descriptor (Randić index of order 0), and three electrostatic descriptors which reflect the bonding properties of molecules, *i.e.* their capabilities to create (mainly) hydrogen bonds. Because of that, hydrogen-donors and hydrogen-acceptors surface areas, charges, total molecular surface areas, and maximum net atomic charges and state energies for oxygen atoms appear to be key factors for modeling the viscosity of organic molecules.

Keywords : Viscosity, molecular structure, correlation, QSPR.

Introduction

Developing models for predicting different not easily measurable experimental properties of molecules is a growing field of research. Progress in this field is accelerated by an enormous increase of computer power, accompanied by development of a number of program packages for graphical and computational manipulation with molecules, as well as for calculation of a huge number of descriptors purely from chemical structure of molecules. Then these easily computable descriptors are used as input information for developing models for computing/predicting values of other not-so-easy-measurable experimental properties of molecules.

Usefulness of molecular modeling methods will be illustrated by modeling the viscosity of liquids, which is one of the most important collective properties of molecules (liquids). Viscosity is one of the most significant transport properties for many chemical engineering problems (*e.g.* petroleum chemistry) and for monitoring (and preventing) spreading of known or not yet synthesized compounds as possible environmental pollutants.[1]

This report represents an extension of our previous studies.[2–4] The data set of molecules is the same as in our previous paper.[3] Molecular descriptors are the same as in ref. 4, and were computed by the CODESSA program.[5] In the model development stage three variable selection methods will be used: (1) heuristic method from version 2.21 of the improved CODESSA program;[5,6] (2) Genetic Function Approximation (GFA) method[7] as implemented in Cerius2 program package,[8] and (3) CROMRsel based approach.[9,10] A short description of these methods is given in Experimental section. Here we will compare these three methods for the model generation used in the field of QSPR (quantitative structure-property relationship) in order to rank them according to their usefulness in developing: (1) good (stable and predictable) QSPR models, and (2) facile/straightforward (easily interpretable) models. Some of the descriptor selection methods used in modeling process (like Neural Network Ensemble (NNE))[11] produce models that are not simple and that are difficult to interpret. We have shown in our recent study that such models can be replaced by simpler multiregression (MR) models,[12] and due to that fact we concentrate here on multiregression based methods. Nonlinear viscosity models will be generated only using the CROMRsel approach. Finally, methods used will be strictly compared (under the same conditions) and models obtained will be critically analyzed from a viewpoint of their goodness and simplicity.

Results and Discussion

In this study, experimental viscosities of 361 structurally diverse organic compounds containing C, H, O, N, S, were used. Each of the three methods (CODESSA, GFA and CROMRsel) was applied to this set of data. Up to 822 molecular descriptors were calculated using the CODESSA program encoding properties of each of the 361 molecules,[4–6] *i.e* . the

same set of descriptors as used in ref. 4. In the CODESSA program, descriptors were computed in a totally automated way solely from the chemical structures of compounds. CODESSA can generate a large number of quantitative descriptors that encode constitutional, topological, geometrical, electrostatic, and quantum chemical characteristics of a molecule. After filtering, 420 descriptors remain.

Table 1. The best two- to eight-descriptor linear multiregression viscosity (log η) models of 361 compounds selected from 420 descriptors by CODESSA (heuristic method from version 2.21)

Model[a]	descriptors[b]	R^{2} [c]	R_{cv}^{2}[c]	S[d]	S_{cv}[d]
MCOD-2	d19, d117	0.8150	0.8091	0.249	0.253
MCOD-3	D19, d117, d218	0.8310	0.8248	0.238	0.242
MCOD-4	d16, d20, d96, d117	0.8433	0.8359	0.229	0.234
MCOD-5	d16, d20, d99, d117, d230	0.8536	0.8446	0.221	0.228
MCOD-6	d20, d117, d134, d143, d154, d371	0.8639	0.8548	0.213	0.220
MCOD-7	d16, d20, d99, d117, d156, d226, d280	0.8717	0.8616	0.207	0.215
MCOD-8	d19, d128, d211, d226, d280, d305, d344, d393	0.8763	0.8670	0.203	0.211

[a] *e.g.* MCOD-2 is abbreviation related to the model (M) obtained by CODESSA (COD) containing two (-2) descriptors; [b] descriptors are designated according to their order in data set (*i.e.* d1, …, d420) – the meaning of selected descriptors is given in Table 7; [c] square of the fitted (R) and the LOO CV (R_{cv}) correlation coefficient; [d] fitted (S) and LOO CV (S_{cv}) standard error of estimate having $N = 361$ in denominator. Please note that S and S_{cv} in ref. 4 were computed using $N - I - 1$ in denominator, where I stands for the number of descriptors involved in the model.

The best multiregression models obtained by the standard CODESSA (version 2.21) descriptor selection procedure containing two to seven descriptors were generated. Details about these models and their statistical parameters are given in Table 1. Description of heuristic selection procedure used in CODESSA for selection of models is given in Experimental Section. Quality of models is expressed by computing the correlation coefficient and standard error of estimate both in the fitting (R and S) and in the leave-one-out (LOO) cross-validation (CV) procedure(s) (R_{cv} and S_{cv}). For detailed description of LOO CV procedure see Experimental section or refs. 9 and 10. Multiregression equation, *i.e.* regression coefficients, their errors and *t*-test values, corresponding to the five-descriptor CODESSA model MCOD-5 from Table 1 is given in Table 2. This model is the same as the five-descriptor CODESSA model from ref. 4.

Table 2. The best five-descriptor model (MCOD-5) selected by CODESSA

Descriptor	short name	$X \pm \Delta X^a$	t-test^b
	intercept	–10.3 ± 1.7	–6.1
d16	Rel. N_{rings}	(2.78 ± 0.41)	6.8
d20	G_I(all pairs)	(55.7 ± 1.6)x10^{-5}	34.2
d99	FPSA(3)	20.2 ± 3.1	6.6
d117	HA dep. HDCA	1.77 ± 0.08	23.0
d230	E_{min}(C)	(89.7 ± 16.5) x10^{-3}	5.4
stat. parameters: R^2 = 0.8536, R_{cv}^2 = 0.8446, S = 0.221, S_{cv} = 0.228, F = 414.1			

[a] X = regression coefficient, ΔX= error of regression coefficient; [b] note: descriptors having higher t-test value are more significant ones

In Table 3 the best MR models containing one to five descriptors selected by the GFA method, which is incorporated in Cerius2 program package, are given. The standard GFA procedure was used with default initial parameters as suggested in Cerius2 package. Description of the GFA selection procedure is given in Experimental Section. Details about the best five-descriptor viscosity model selected by the GFA method are given in Table 4.

Table 3. The best one- to five-descriptor linear multiregression viscosity models of 361 compounds selected from 420 descriptors by the GFA method (included in Cerius2 package – ref. 8)

Model[a]	descriptors[b]	$R^{2\,c}$	R_{cv}^{2c}	S^d	$S_{cv}^{\ d}$
MGFA-1	d92	0.3624	0.3558	0.462	0.464
MGFA-2	d20, d117	0.8106	0.8040	0.252	0.256
MGFA-3	d19, d35, d117	0.8331	0.8265	0.236	0.241
MGFA-4	d19, d22, d117, d143	0.8530	0.8466	0.222	0.227
MGFA-5	d19, d22, d116, d119, d211	0.8652	0.8573	0.212	0.218

[a] *e.g.* MGFA-1 is abbreviation related to the model (M) obtained by the GFA (GFA) method containing one (-1) descriptor; [b, c, d] see footnotes in Table 1

Table 4. The best five-descriptor model (MGFA-5) selected by GFA

Descriptor	short name	$X \pm \Delta X^a$	t-test^b
	intercept	–0.78 ± 0.03	–24.1
d19	G_I(all bonds)	(22.5 ± 1.5) x10^{-4}	15.0
d22	$^0\chi$, Randić connect. ind. of order 0	(–131.1 ± 17.4)x10^{-3}	–7.5
d116	HA dep. HDCA-1/TMSA	–146.6 ± 20.0	–7.3
d119	HA dep. HDCA-2/SQRT(TMSA)	78.2 ± 6.0	13.0
d211	$P_{\pi\text{-}\pi}$, max π - π bond order	(–135.8 ± 25.2) x10^{-3}	–5.4
stat. parameters: R^2 = 0.8652, R_{cv}^2 = 0.8573, S = 0.212, S_{cv} = 0.218, F = 455.8			

[a, b] see footnotes in Table 2

The third method used was the CROMRsel procedure for selecting the best possible subsets of descriptors.[9,10] This procedure searches over all possible subsets having I (I = 1, …, 5 in this case) descriptors, and, for each I selects the best model. In this case intercorrelation between descriptors involved in the MR model is minimised. Due to this fact, one can expect that such models have the best statistical performance, *i.e.* fitted, cross-validated and predictive performances. The best one- to five-descriptor models selected by the CROMRsel procedure for selecting the best possible combination (subset) of descriptors are given in Table 5.

Table 5. The best possible one- to five-descriptor linear multiregression viscosity models of 361 compounds selected by CROMRsel from 420 descriptors

Model[a]	descriptors[b]	$R^{2\,c}$	R_{cv}^{2c}	S^d	$S_{cv}{}^d$
M-1	d96	0.4852	0.4770	0.415	0.418
M-2	d19, d117	0.8150	0.8091	0.249	0.253
M-3	d19, d22, d117	0.8396	0.8336	0.232	0.236
M-4	d15, d17, d99, d117	0.8552	0.8483	0.220	0.225
M-5	d19, d22, d115, d304, d306	0.8735	0.8636	0.206	0.214

[a] *e.g.* M-1 is abbreviation related to the model (M) obtained by the CROMRsel selection method containing one (-1) descriptor; [b, c, d] see footnotes in Table 1

Table 6. The best linear five-descriptor model (M-5) selected by CROMRsel

Descriptor	short name	$X \pm \Delta X^a$	t-test[b]
	intercept	$(-840.5 \pm 28.8)\times10^{-3}$	–29.2
d19	G_I(all bonds)	$(22.6 \pm 1.4)\times10^{-4}$	16.2
d22	${}^0\chi$, Randić connect. ind. of order 0	$(-132.0 \pm 16.8)\times10^{-3}$	–7.9
d115	HA dep. HDCA-1	$(265.1 \pm 28.8)\times10^{-3}$	9.2
d304	HA dep. HDCA-1/TMSA	–138.1 ± 12.6	–11.0
d306	HA dep. HDCA-2/TMSA	775.4 ± 65.1	11.9
stat. parameters: R^2 = 0.8735, $R_{cv}{}^2$ = 0.8636, S = 0.206, S_{cv} = 0.214, F = 490.2			

[a, b] see footnotes in Table 2

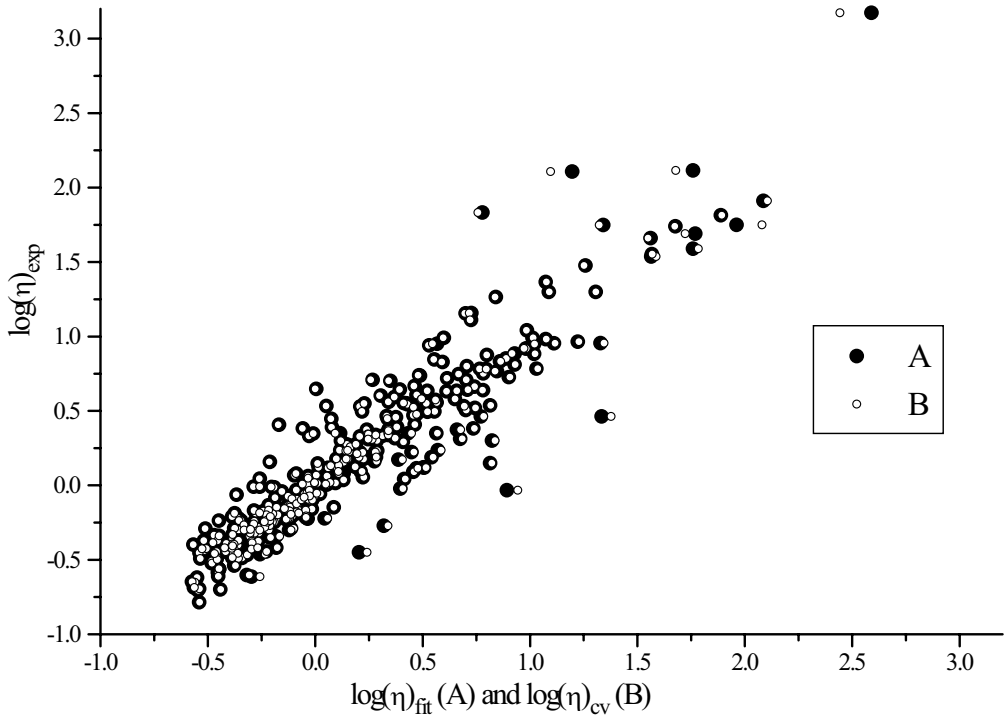

Figure 1. Scatter plot of the calculated (fit (A) and LOO cross-validated (B)) *vs* experimental log η for 361 compounds using the linear model from Table 6.

Table 6 gives additional details and statistical parameters of the best five-descriptor model obtained by CROMRsel. All descriptors involved in all CODESSA, GFA and CROMRsel models in Tables 1, 3 and 5 are listed and described in Table 7. Additional explanation of descriptors can be found in refs. 13 and 14. Calculated values for 361 compounds by the fit and LOO CV procedures *versus* experimental viscosities using model M-5 from Table 6 are given in Figure 1.

Table 7. Molecular descriptors involved in the MR viscosity models selected in this study [a]

No.	descriptors
d15	N_{rings}, Number of rings
d16	Relative number of rings
d17	Molecular weight
d19	G_I, Gravitation index (all bonds)
d20	G_I, Gravitation index (all pairs)
d22	$^0\chi$, Randić index (order 0)
d35	Information content (order 0)
d77	Polarity parameter (Q_{max} - Q_{min}) [Zefirov's PC]
d80	Topographic electronic index (all bonds) [Zefirov's PC]
d92	FPSA-2 Fractional PPSA (PPSA-2/TMSA) [Zefirov's PC]
d96	PPSA-3 Atomic charge weighted PPSA [Zefirov's PC]
d99	FPSA-3 Fractional PPSA (PPSA-3/TMSA) [Zefirov's PC]
d115	HA dependent HDCA-1 [Zefirov's PC]
d116	HA dependent HDCA-1/TMSA [Zefirov's PC]
d117	HA dependent HDCA-2 [Zefirov's PC]
d119	HA dependent HDCA-2/SQRT(TMSA) [Zefirov's PC]
d128	HACA-2/TMSA [Zefirov's PC]
d134	HOMO-1 energy
d143	Maximum electrophilic reactivity index for a C atom
d144	Average electrophilic reactivity index for a C atom
d154	Total point-charge component of the molecular dipole
d156	Total dipole of the molecule
d211	$P_{\pi\text{-}\pi}$, Max π-π bond order
d218	Maximum bond order of a C atom
d226	Minimum electron-electron repulsion for a C atom
d230	E_{min}(C), Minimum atomic state energy for a C atom
d280	Principal moment of inertia C
d295	Minimum(no. of HA, no. of HD) [**Semi-MO PC**]
d299	HA dependent HDSA-1/TMSA [**Semi-MO PC**]
d302	HA dependent HDSA-2/SQRT(TMSA) [**Semi-MO PC**]
d304	HA dependent HDCA-1/TMSA [**Semi-MO PC**]
d305	HA dependent HDCA-2 [**Semi-MO PC**]
d306	HA dependent HDCA-2/TMSA [**Semi-MO PC**]
d344	Positively charged SA
d371	PCSA-2 of C atoms
d393	Q_{max}(O), Maximum net atomic charge for a O atom
d406	E_{max}(O), Max atomic state energy for a O atom

[a] Numbering of descriptors corresponds to that in the 420-descriptor data set; TMSA = total molecular surface area; PPSA = partial positive surface area; HDSA = hydrogen-bonding donor surface area; FPSA-3 = FPSA-3 fractional PPSA (PPSA-3/TMSA); PPSA-3 = PPSA-3

atomic charge weighted PPSA; FPSA-2= fractional PPSA (PPSA-2/TMSA); PPSA-2 = total charge weighted PPSA; HA = hydrogen-acceptors; HD = hydrogen-donors; HDCA = hydrogen-donors charged surface area; HACA = hydrogen-acceptors charged surface area; HACA-2 = total charge weighted HACA; HOMO = the highest occupied molecular orbital; PCSA = positively charged surface area; vdW = van der Waals radius; SA = surface area.

Finally, nonlinear modeling using CROMRsel algorithms was performed. Starting from initial 420 descriptors 25 descriptors were preselected by using stepwise CROMRsel procedure called SSP2 (for description see Experimental section and ref. 10). Then, nonlinear terms in the form of squares and cross-products of initial 25 descriptors were computed and added to the initial 25 descriptors. Data set created in such a way contained 350 descriptors and was a starting set for generating nonlinear models. Using the standard CROMRsel procedure for the selection of the best possible subset of descriptors, the best one- to five-descriptor models were obtained and details of the best selected models are given in Table 8. One can see that models containing two to five descriptors include initial (linear) descriptors d19 and d22 (except Mnonlin-2 model), as well as cross-products of initial descriptors. In this study only two-fold cross-products were used. The value of a two-fold cross-product descriptor for a molecule is just the product of initial (single) descriptor values of two descriptors calculated for the molecule. Calculated values for 361 compounds by the fit and LOO CV procedures *versus* experimental viscosities using model Mnonlin-5 from Table 9 are given in Figure 2.

Table 8. The best possible one- to five-descriptor nonlinear multiregression viscosity models of 361 compounds selected by CROMRsel from 420 descriptors

Model[a]	descriptors[b]	$R^{2\,c}$	R_{cv}^{2c}	S^{d}	S_{cv}^{d}
Mnonlin-1	d80 x d99	0.5663	0.5597	0.381	0.384
Mnonlin-2	d19, d77 x d295	0.8428	0.8380	0.229	0.233
Mnonlin-3	d19, d22, d295 x d393	0.8696	0.8652	0.209	0.212
Mnonlin-4	d19, d22, d295 x d295, d295 x d393	0.8878	0.8837	0.194	0.197
Mnonlin-5	d19, d22, d99 x d302, d299 x d406, d302 x d393	0.9082	0.9036	0.175	0.180

[a] *e.g.* Mnonlin-1 is abbreviation related to the model (M) which include nonlinear terms (nonlin) obtained by the CROMRsel method containing one (-1) descriptor; [b] *e.g.* d80 x d99 denotes cross-product of descriptors d80 and d99 (see also footnote (b) in Table 1); [c, d] see footnotes in Table 1

Table 9. The best nonlinear five-descriptor viscosity model (Mnonlin-5) selected by CROMRsel

Descriptor	short name	$X \pm \Delta X^a$	t-test^b
	intercept	$(-755.8 \pm 24.4)\times10^{-3}$	–31.0
d19	G_I (all bonds)	$(23.3 \pm 1.2)\times10^{-4}$	19.5
d22	$^0\chi$ (Randić connect. ind. of order 0)	$(-146.8 \pm 14.4)\times10^{-3}$	–10.2
d99 x d302	(PPSA-3/TMSA) x (HDSA-2/sqrt(TMSA)	47.5 ± 5.1	9.4
d299 x d406	(HDSA-1/TMSA) x E_{max}(O)	$(-69.7 \pm 4.7)\times10^{-3}$	–14.7
d302 x d393	(HDSA-2/sqrt(TMSA) x Q_{max}(O)	–25.6 ± 1.4	18.0
stat. parameters: $R^2 = 0.9082$, $R_{cv}^2 = 0.9036$, S = 0.175, $S_{cv} = 0.180$, $F = 702.2$			

$^{a, b}$ see footnotes in Table 2

It is evident from Figures 1 and 2 that the difference between corresponding fitted and LOO CV values is much smaller in the case of the best nonlinear five-descriptor model, indicating model stability. Stability of this five-descriptor model is additionally tested by performing partition of 361 molecules into the set containing molecules 1 - 240 from Table 1 in ref. 4 (training set, *i.e.* set on which model was generated) and the external set containing remaining 121 molecules (241- 361) for which viscosities were calculated by the model generated on the training set. Standard errors of estimate of several top models on the training and external sets given in Table 10 show that the models selected in this study are good and stable. Linear models having the same number of descriptors selected by the CROMRsel method are somewhat better in prediction on external set than corresponding models obtained by CODESSA and GFA. Nonlinear models obtained by the CROMRsel procedures are considerably better than the linear ones. The best nonlinear five-descriptor model has the standard error of 0.158 log units on external set, and linear five-descriptor models obtained by CODESSA, GFA and CROMRsel have standard error of about 0.21 log units (Table 10). Moreover, the best nonlinear two-descriptor model has smaller standard error of estimate on external set (0.198 lo units) than the best linear five-descriptor models.

Table 10. The stability test of the best selected models by performing an external validation

Model[a]	***S*** (log units)	
	trening set: 240 molecules	external set: 121 molecules
MCOD-3	0.248	0.218
MCOD-4	0.234	0.223
MCOD-5	0.230	0.210
MCOD-8	0.209	0.202
MGFA-2	0.256	0.247
MGFA-3	0.238	0.238
MGFA-4	0.230	0.208
MGFA-5	0.215	0.211
M-2	0.258	0.232
M-3	0.238	0.222
M-4	0.230	0.202
M-5	0.205	0.209
Mnonlin-2	0.244	0.198
Mnonlin-3	0.224	0.178
Mnonlin-4	0.207	0.167
Mnonlin-5	0.184	0.158

[a] Abbreviations are related to the models given in Tables 1, 3, 5 and 8.

From the statistical point of view, one can conclude that all the models selected in this study are highly significant and stable. This can be best seen from the calculated standard errors obtained on external set of molecules (Table 10) and from the differences between the fitted (S) and LOO cross-validated (S_{cv}) standard errors of estimate. Additionally, significance of each descriptor involved in the best five-descriptor models selected in this study can be seen in Tables 2, 4, 6 and 9 through their *t*-test values. However, one can say that in the case of linear models the CROMRsel method generates better models. The CROMRsel procedure is simple for use because only the size of models is input, which is defined by the number of descriptors that should be involved in the model. The use of CODESSA and GFA methods is also simple, but in these two methods there is at least one parameter more (than in the case of CROMRsel) that should be given or optimised by the user (see the description of selection methods used in GFA and CODESSA 2.21 in Experimental section). In the next paragraph, an example related to the optimization of the significant intercorrelation level between descriptors that can be involved in the CODESSA models is given. On the other hand the CROMRsel procedure is more time consuming (comparing with GFA and CODESSA) in the case of models containing more descriptors. In this study the selection of four- and five-descriptor models used more computer time than the other two methods.

For this data set, models obtained by CROMRsel are the best possible one- to five-descriptor multivariate regression models. Therefore, these models can be used for finding out better input parameters that should be defined by users before running variable selection procedures in the CODESSA and GFA methods. After changing default value of the significant intercorrelation level in CODESSA 2.21 the best five-descriptor model was improved from R^2 = 0.8536 (R_{cv} = 0.8446, F = 414.1) to R^2 = 0.8615 (R_{cv} = 0.8536, F = 441.8). According to the statistical parameters this model is almost as good as the five-descriptor CROMRsel and GFA models. This test indicates that some input parameters in CODESSA (like an intercorrelation level) should be softened, *i.e.* the selection procedure should be repeated with different values of input parameters in order to improve models. Results presented here are obtained with default input parameters in order to test these three methods '*as they are originally given*' without additional optimizations. In the GFA method only default initial parameters were used because it is not easy to know in advance what is the best range of input parameters (like the number of generations and the value of smoothing parameter).

Nonlinear models (having the same number of descriptors and optimized parameters as linear models) generated by the use of CROMRsel algorithms are better than linear ones. An improvement obtained by the best four- and five-descriptor nonlinear models (R_{cv}^2 = 0.884 and R_{cv}^2 = 0.904, respectively) with respect to the previous linear model obtained on the same set of data[4] is evident (the best four- and five-descriptor models in ref. 4 had R_{cv}^2 = 0.852 and R_{cv}^2 = 0.844, respectively).

By comparing the type of descriptors involved in the five-descriptors models one can see that each models contains mainly two type of descriptors: (1) geometrical and/or topological descriptors, and (2) several (usually three) electrostatic descriptors which describe the potency of molecule for forming preferably hydrogen bonds.

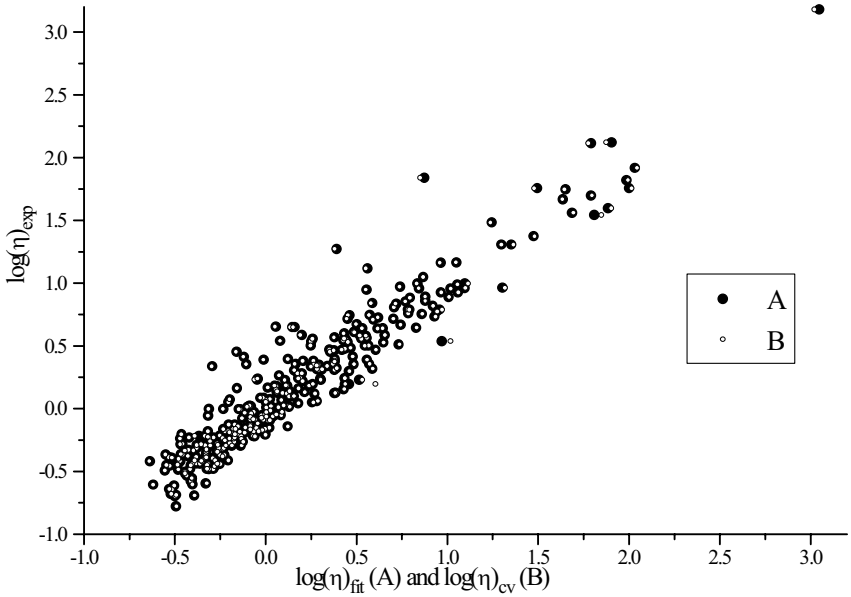

Figure 2. Scatter plot of the calculated (fit (A) and LOO cross-validated (B)) *vs* experimental log η for 361 compounds using the nonlinear model from Table 9.

From the first class of descriptors, the gravitational index is involved in each model. In addition, in all five-descriptor models this descriptor is the most significant one (*i.e.* it has the highest *t*-test value). The gravitational index reflects the effective mass distribution in the molecule and effectively describes intermolecular dispersion forces in the bulk liquid media (*i.e.* accounts simultaneously for both the atomic masses and for their distribution within the molecular space).[4] The second descriptor from the first class is the Randić index[15] of order 0, which is related to the number of atoms in molecule. In one case (the CODESSA five-descriptor model) instead of the Randić index a descriptor related to the relative number of rings in the molecule is involved. This descriptor accounts for both the size and shape of a molecule.

From the second class of descriptors the most often involved descriptors are those related to the hydrogen-donor(s) and hydrogen-acceptor(s) charged surface areas of molecules, total molecular surface areas, and maximum net atomic charges and state energies for oxygen atoms in organic molecules. These descriptors are related to the hydrogen bonding ability of compounds and to the ability of forming other polar interactions between solute molecules. This reflects that the mass, size, shape as well as the hydrogen bonding ablities of molecules are key factors which govern their liquid viscosity.

Experimental Section

Data set. The list of 361 molecules was taken from ref. 3. Complete list of molecules (names), experimental viscosity values and viscosities computed using the best linear and non-linear five-descriptor models from Tables 2, 4, 6 and 9 can be obtained on request. The range of experimental viscosity values of 361 molecules is between 0.164 (trans-2-pentene) and 1490 (glycerol) mPa·s (measured at 20 °C). In this study, the logarithmically transformed ($\log_{10}$) viscosity values were used.

Calculation of descriptors. The procedure for computing molecular descriptors by the CODESSA program is described in ref. 4. By using the CODESSA program, the number of descriptors with up 822 for 361 diverse organic molecules were calculated. However, from the total number of 822 descriptors any descriptor containing more than 80% zeros or identical values was eliminated. By this filtering procedure 402 descriptors were removed, and the set of 420 descriptors remained and was used in the modeling. Zero values were assumed for missing values of each descriptor**.** In addition, the list of 822 descriptors (and, also, the list of 420 descriptors which were obtained after filtering) and their values for 361 molecules can also be obtained in electronic form on request.

Heuristic descriptor selection method.[4,5] In the case of the heuristic method used in CODESSA, the selection of descriptors for multivariate regression models was performed based on the several criteria. The algorithm used several parameters (options), which control the performance of the method. Default values of the parameters were selected to fit the most common situations (these parameters can be easily changed in the respective dialog box). Firstly, descriptors from the "starting set" of descriptors were eliminated if: (a) the *F*-test's

value for the one-descriptor correlation with the descriptor was below 1.0, (b) the squared correlation coefficient of the one-descriptor model was less than R_{min}^2, (c) the descriptor's *t*-value was less than t_1 (0.1), (d) the descriptor was highly intercorrelated above r_{full} with another descriptor. The values of R_{min}^2, t_1 and r_{full} were user specified. In this study the default values of these parameters were used for the models given in Table 1, *i.e.* 0.01, 0.1 and 0.99 for R_{min}^2, t_1 and r_{full}, respectively. Then the following procedure followed:

(1) Starting with the top descriptor from the preselected list of descriptors all two-descriptor models were calculated for pair of descriptors having intercorrelation coefficient lower than $r_{sig} = 0.8$.
(2) The best 10 two-descriptor models were selected and processed further to the stepwise selection procedure for the development of multidescriptor models.
(3) To the multivariate regression models containing *n* descriptors a new descriptor (having the intercorrelation coefficient with other involved descriptors lower than 0.8) was added to generate a model with *n* +1 descriptors.
(4) The best 10 (*n*+1)-descriptor models were again submitted to the same procedure, until the multivariate regression model with a certain number of descriptors was obtained.

Genetic Function Approximation (GFA) method.[7] The GFA algorithm uses a genetic algorithm to perform a search over the space of possible QSAR/QSPR models using the LOF score as a parameter for estimating the fitness of each model. Such evolution of a population of randomly constructed models leads to the discovery of highly predictive QSARs/QSPRs. Genetic algorithms were derived by analogy with the spread of mutations in a population. According to this analogy "individuals" are represented as a 1D string of bits. An initial population of individuals is created, usually with random initial bits. In the GFA method, models containing a randomly chosen proper subset of the independent variables are collected and then the collected models are "evolved". A *generation* is the set of models resulting from performing the multiple linear regression on each model. A selection of the best ones becomes the next generation (set of models). Crossover operations are performed on these, which take some variables from each of two models to produce an offspring. In addition, the best model from the previous generation is retained. Besides linear terms, there can also be spline, quadratic and quadratic spline terms. These are added or deleted by mutation operations. A disadvantage is that it is not possible to introduce cross-products of descriptors as nonlinear terms in the GFA method included in Cerius2. Only default values of input parameters were used in GFA (the default values for the number of generations and smoothing parameter were 5000 and 1.0, respectively).

CROMRsel procedure.[9,10] For generating linear CROMRsel models the algorithm for selection of the possible subsets of descriptors was used. Detailed description of this algorithm was given in ref. 9 and the final result, which one can obtain by using this CROMRsel algorithm, is selection of the best possible models containing subsets of *I* descriptors ($I = 1, \ldots, 5$ in this study).

SSP2 CROMRsel procedure.[9,10] For generating nonlinear CROMRsel models, the preselection of descriptors was performed by the Stepwise Selection Procedure denoted as no. 2 in ref. 10 (SSP2). SSP2 selects up to *K* descriptors ($K = 25$ in this study) in the '*1 by 1*'

manner (adding one new descriptor to the model in each step). In this procedure the stepwise selection starts from each (j) descriptor (j = 1, 2, ..., N; N = 420 in this study) in the data set, giving N models each with K descriptors. Among them the best one (according to the highest fit correlation coefficient) is chosen. Then, starting from 25 descriptors preselected by SSP2 procedure and their squares and cross-products (350 descriptors altogether), the standard CROMRsel algorithm for selection of the possible subsets of descriptors was used in order to obtain nonlinear models given in Table 8.

Cross-validation. LOO CV procedure is a procedure usually used for evaluating the model stability. During LOO CV procedure each of N molecules is taken away only once. Using the remaining N-1 molecules, multiregression models were generated and using that model in each step viscosity values for excluded compound are calculated. Finally, we have prediction (in LOO CV sense) of viscosity values for N molecules.

Cross-product of descriptors. From purely mathematical point of view, the cross-product of two descriptors is the simplest form of a nonlinear function one can generate from the two initial descriptors. For example, descriptor d302 x d393 indicates existence of nonlinear dependencies (in functional form of a cross-product) between descriptors d302 and d393 on one side and on the other side experimental viscosity values for 361 molecules analysed in this study.

Acknowledgements

The authors thank reviewers for their helpful comments.

References and Notes

1) Reid, R. C.; Prausnitz, J. M.; Poling, B. E. *Properties of Gases and Liquids*, 4th ed. McGraw-Hill, New York (1987).
2) Suzuki, T.; Ohtaguchi, K.; Koide, K. *Comput. Chem. Eng*. **1996**, *20*, 161.
3) Suzuki, T.; Ebert, R.-U.; Schüürmann, G. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 1122.
4) Katritzky, A. R.; Chen, K.; Wang, Y.; Karelson, M.; Lučić, B.; Trinajstić, N.; Suzuki, T.; Schüürmann, G. *J. Phys. Org. Chem*. **2000**, *13*, 80.
5) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. CODESSA Version 2.0 Reference Manual, 1994.
6) Katritzky, A. R.; Tatham, D. B.; Maran, U. *J. Chem. Inf. Comput. Sci*. **2001**, 41, 1162.
7) Rogers, D.; Hopfinger A. J. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 854.
8) Cerius2; Accelrys: 9685 Scranton Road, San Diego, CA, 92191.
9) Lučić, B.; Trinajstić, N. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121; Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 610.
10) Lučić, B.; Amić, D.; Trinajstić, N.. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 403.
11) Kovalishyn, V. V.; Tetko, I. V.; Alessandro Villa, A. E. P.; Livingston, D. J. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 651.

12) Lučić, B.; Nadramija, D.; Bašic, I; Trinajstić, N. in preparation.
13) Karelson, M. *Molecular Descriptors in QSAR/QSPR*, Wiley-Interscience, John Wiley & Sons, New York (2000).
14) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev*. **1996**, *96*, 1027.
15) Randić, M. *J. Am. Chem. Soc*. **1975**, *97*, 6609.