# A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors

Bono Lučić* and Nenad Trinajstić

The Rugjer Bošković Institute, P.O. Box 1016, HR-10001 Zagreb, Croatia

Sulev Sild,[†,‡] Mati Karelson,[‡] and Alan R. Katritzky*[,†]

Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, P.O. Box 117200, Gainesville, Florida 32611-7200, and Department of Chemistry, University of Tartu, Jakobi Street 2, EE 2400 Tartu, Estonia

Received November 11, 1998

The selection of the most relevant variable is a frequent problem in the analysis of chemical data, especially now considering the large amounts of data created by the increased computer power and analytical resolution. A novel procedure for variable selection based on multiregression (MR) analysis is developed and applied to the quantitative structure−property relationship (QSPR) modeling of gas chromatographic retention times $t_R$ and Dietz response factors RF on 152 diverse chemical compounds. Using 296 descriptors generated by the CODESSA program, "absolutely the best" linear MR models containing from 1 to 5 descriptors were first selected ($\sim 2 \times 10^{10}$ models were checked), and then "the best" linear stepwise MR models with six and seven descriptors were obtained through "*i* by *i*" stepwise selection. In this paper *i* was varied from 1 to 4, so that in each next step *i* descriptors were added to the previously selected descriptors. Nonlinear models were developed by the inclusion of cross-products of initial descriptors. We selected as the most important descriptors for $t_R$ the number of C−H and C−X bonds, connectivity indices of order 3, the highest normal mode vibrational frequency, and the rotational entropy of the molecule at 300 K. In the case of RF modeling the most important descriptors are those related to the relative number and weight of effective C atoms, the orbital electronic population, and the bond order and valency of C and H atoms. Comparison with the best six-descriptor models obtained by the normal CODESSA procedure shows that nonlinear seven-descriptor MR models now obtained achieve 30% (0.3520 vs 0.5032) and 12% (0.0472 vs 0.0530) less standard errors of estimate for $t_R$ and RF, respectively. Our novel procedure of selecting a small number of the most important descriptors from a data set allows us to extract a larger amount of useful information than with the procedure implemented in CODESSA. Thus, our new procedure enables the selection of the best possible MR models from $10^{10}$ possibilities. Through the introduction of cross-product terms, we obtained nonlinear MR models which are superior to the corresponding linear models.

## INTRODUCTION

The most important aim of mathematical and statistical methods in chemistry is to provide the maximum information about selected molecular property by analyzing chemical data. The quality of a method is reflected in its ability to extract the most relevant information starting from a standard data set. In the case of quantitative structure−property or structure−activity modeling on a given set of molecules, sometimes a large number of descriptors is produced in the first step. Nowadays, there are computer programs available by which one can generate several hundreds of descriptors for modestly sized molecules. Among them, the ADAPT[1−7] and CODESSA programs[8−19] are often used. The use of these programs in structure−property modeling, very often, results in more descriptors than the number of molecules in the data set. A much studied problem is how many descriptors should be used in the final model.[20] A related question (which stems from Ockham's Razor;[21] to prefer the model realized with the fewest descriptors, other things being equal) is how to select a small number of the most important descriptors from a large data set.

Generally, the selection of a small number of important descriptors from a large initial set can be carried out by selecting the relatively small number of descriptors that contain the maximum retention time mapping information (for such a method we used the term "inductive"), reducing the total pool of descriptors by removing those that contain the maximum amount of redundant information (called "deductive" method), or using a combination of the deductive and inductive methods. Most popular program packages solve the problem of selection of the most important descriptors inductively, in a stepwise manner, selecting one descriptor at a time ("one by one" stepwise selection).[8] However, several recently published algorithms for variable selection are essentially deductive[22,23] or utilize a combination of the deductive and inductive approaches.[8,24,25] One of the these is CODESSA.[8] This program can generate a large number

[†] University of Florida.
[‡] University of Tartu.

10.1021/ci980161a CCC: $18.00 
Published on Web 03/25/1999

of descriptors, as well as obtain a structure−property or structure−activity correlation (multiregression (MR) model). In the CODESSA treatment, descriptors that were correlated with other descriptors less than some defined amount were selected in each step. A major limitation of the CODESSA modeling subroutine, in common with other approaches for variable selection, is, for example, the impossibility of selecting the best possible two or three descriptors for MR models from a data set containing, for example, 200 descriptors, for which case the numbers of possible two- and three-descriptor models are 4950 and 161 700, respectively. This problem has been solved rigorously in several cases but, until now, only for relatively small data sets.[26−29]

We now test the possibilities and the quality of a new procedure for descriptor selection on the large data set of descriptors generated by CODESSA.[8] As a consequence of the assumption of the validity of Ockham's Razor,[21] this procedure starts with the selection of the most important descriptors in the multiregression models. Therefore, whenever it was possible (regarding the total number of possible MR models with $I$ descriptors that can be obtained from the total set of $N$ descriptors), the best possible MR models with $I$ descriptors are selected. In other cases this procedure is applied after a preselection procedure which is not the best for each step, but is the best for adding $i$ ($i = 1, ..., 4$) descriptors at a time ("$i$ by $i$"). Such a procedure gives the best stepwise (but not the absolutely best possible) MR model. Descriptors selected in this manner should contain more information than in the case of "*one* by *one*" selection. The efficient selection among such a large number of descriptors (and, consequently, a large number of models) was made possible by working with orthogonal descriptors. One of the aims of this paper is to show that one can obtain by our procedure better models with the use of the standard CODESSA selection procedure.[8−19] Moreover, the final descriptors selected for linear MR models were used for the calculation of cross-product terms and, hence, for obtaining nonlinear MR models. Whenever the new data set of both linear (initial) and nonlinear (cross-products) descriptors was to large to allow the selection of the best possible models with $I$ (for example, $I = 7$) descriptors, the stepwise ($i$ by $i$) procedure was applied.

Experimentally, the response factors and especially retention times can be measured very accurately. Therefore, the models developed for the prediction of these properties should have prediction errors as small as the experimental errors, if possible. If the experimental procedure can separate two molecules (i.e., can measure the difference in their retention times), it would be highly desirable that the models can do the same, i.e., to predict their separate retention times. Of course, this ultimate goal will be possible only for pairs of molecules whose actual retention times are different for more than the prediction error of a model. From this consideration it follows that it is important to search for models that will be more accurate than the models obtained by the CODESSA,[8] i.e., to make efforts for developing such a procedure by which it will be possible to extract the most representative information from a descriptor set, which is needed for accurate prediction.

## EXPERIMENTAL SECTION

The computations were done on a Hewlett-Packard 9000/E55 (PA-RISC 7100LC processor, 100 MHz) and SUN Enterprise 3000 (UltraSPARC processor, 250 MHz) in multiuser mode. The FORTRAN 77 computer programs were developed for the selection of (1) the best possible multiregression models and (2) the best stepwise multiregression models ("the best" means the best according to the correlation coefficient). These programs include original subroutines for orthogonalization of descriptors and for cross-validation, as well as the SVDFIT subroutine from Numerical Recipes.[30] The SVDFIT subroutine was used only for the computation of final models based on the best selected descriptors.

**Data Sets.** The data set used is the same as in the Katritzky et al. paper.[8] By the use of the CODESSA program[8,9] 388 descriptors for 152 very diverse and randomly selected organic molecules were generated. However, the number of descriptors with calculated values for each of 152 molecules was reduced to 296. Initially, the modeling began with the maximum number of 388 descriptors, setting 0.0 for those values of descriptors that could not be calculated. This gave no significant improvement. For logical consistency (the impossibility of calculating the descriptor value for a molecule does not necessarily mean that the value of the descriptor is zero), the modeling was continued with 296 descriptors. Of these, 144 were quantum-chemical and 152 conventional descriptors. The conventional set contained *topological indices*, *geometrical*, *informational*, and *structural* descriptors. Structural descriptors were absolute or relative counts of different atoms, bonds, or rings, as well as gravitation descriptors or molecular weight.[8,31] Quantum-chemical descriptors used by CODESSA were computed using the AM1 method[32] (in the framework of the MOPAC program).[33] In the paper of Katritzky et al,[8] the descriptors were divided into conventional and quantum-chemical sets for which preselection of descriptors was done separately. Contrary to the QSPR (quantitative structure−property relationship) treatment done by the CODESSA program, all of the descriptors in our present procedure were treated as equally good, i.e., the set of descriptors was not divided into subsets before the selection procedure.

As described by Katritzky et al,[8] the values of gas chromatographic (GC) retention times ($t_R$) and response factors (RF) for all studied compounds were determined by a Hewlett-Packard 5890 series II gas chromatograph equipped with a flame ionization detector (FID) and split injection port. They are given in Tables 1 and 2 in ref 8, as well as in Table 1 of this paper. All data and related programs needed for the modeling described in this paper are available on request (lucic@faust.irb.hr).

**Multivariate Regression.** A regression equation estimates a dependent variable $P^{est}$ (which is an approximation to an experimentally determined chemical property, $P$) from independent variables called descriptors ($d_i$, $i = 1, ..., I$). This relationship is expressed by a linear combination of descriptors such as

$$P^{est} = (c_0 \pm \Delta c_0) + (c_1 \pm \Delta c_1)d_1 + (c_2 \pm \Delta c_2)d_2 + ... + (c_i \pm \Delta c_i)d_i + ... \quad (1)$$

The coefficients of contribution $c_i$ of each single descriptor

**Table 1.** Comparison of Observed with Fitted (calc) and 1-Fold Cross-Validated ($calc_{cv}$) Predicted Values Calculated by Seven-Descriptor Models (Tables 5 and 7) for Retention Times ($t_R$/min) and Response Factors ($RF_{Dietz}$) and Percentage Deviations of Fitted Values (Δ%) for Studied Compounds

| | | $t_R$ (min) | | | $RF_{Dietz}$ | | | | | $t_R$ (min) | | | $RF_{Dietz}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| no. | compound | obsd | calc (Δ%) | $calc_{cv}$ | obsd | calc (Δ%) | $calc_{cv}$ | no. | compound | obsd | calc (Δ%) | $calc_{cv}$ | obsd | calc (Δ%) | $calc_{cv}$ |
| 1 | dodecane | 6.34 | 6.16 (3) | 6.15 | 0.96 | 0.91 (6) | 0.90 | 76 | indole | 6.81 | 7.02 (3) | 7.03 | 0.61 | 0.67 (10) | 0.68 |
| 2 | pentadecane | 8.56 | 8.43 (1) | 8.41 | 0.90 | 0.83 (7) | 0.83 | 77 | nitrobenzene | 5.09 | 5.23 (3) | 5.26 | 0.63 | 0.59 (6) | 0.59 |
| 3 | 1-octene | 2.53 | 2.47 (2) | 2.46 | 1.00 | 0.99 (1) | 0.99 | 78 | benzothiazole | 6.33 | 6.10 (4) | 6.09 | 0.63 | 0.60 (5) | 0.60 |
| 4 | 1-decene | 4.45 | 4.53 (2) | 4.52 | 0.94 | 0.95 (1) | 0.95 | 79 | benzimidasole | 8.02 | 7.35 (8) | 7.32 | 0.49 | 0.56 (15) | 0.57 |
| 5 | cyclohexanol | 3.27 | 3.64 (11) | 3.66 | 0.80 | 0.74 (8) | 0.73 | 80 | quinaldine | 7.04 | 6.94 (1) | 6.94 | 0.77 | 0.81 (5) | 0.81 |
| 6 | cyclohexanone | 3.25 | 3.24 (0) | 3.24 | 0.76 | 0.77 (1) | 0.77 | 81 | 2,3-lutidine | 3.80 | 4.12 (9) | 4.14 | 0.86 | 0.88 (2) | 0.88 |
| 7 | octylamine | 4.88 | 5.09 (4) | 5.10 | 0.79 | 0.80 (1) | 0.80 | 82 | 2,6-lutidine | 3.23 | 3.49 (8) | 3.50 | 0.86 | 0.88 (3) | 0.88 |
| 8 | decylamine | 6.62 | 6.61 (0) | 6.61 | 0.76 | 0.78 (2) | 0.78 | 83 | phenyl disulfide | 10.15 | 10.67 (5) | 10.77 | 0.54 | 0.52 (4) | 0.52 |
| 9 | hexanenitrile | 3.06 | 2.93 (4) | 2.92 | 0.82 | 0.80 (3) | 0.80 | 84 | 2,4-lutidine | 3.66 | 3.67 (0) | 3.65 | 0.86 | 0.88 (2) | 0.88 |
| 10 | cyclopentanone | 2.26 | 2.26 (0) | 2.26 | 0.72 | 0.76 (5) | 0.76 | 85 | *o*-cresol | 4.84 | 5.03 (4) | 5.03 | 0.82 | 0.77 (6) | 0.77 |
| 11 | dioctyl sulfide | 10.90 | 10.99 (1) | 10.93 | 0.55 | 0.56 (2) | 0.56 | 86 | 1,3,5-trimethoxybenzene | 7.68 | 7.44 (3) | 7.43 | 0.46 | 0.44 (5) | 0.44 |
| 12 | diethyl carbonate | 2.29 | 2.36 (3) | 2.36 | 0.40 | 0.44 (9) | 0.44 | 87 | 4-isopropylphenol | 6.32 | 6.40 (1) | 6.40 | 0.72 | 0.77 (6) | 0.77 |
| 13 | cyclopentanol | 2.32 | 2.93 (26) | 2.95 | 0.73 | 0.73 (0) | 0.73 | 88 | benzophenone | 9.23 | 9.22 (0) | 9.23 | 0.60 | 0.68 (13) | 0.69 |
| 14 | 1-octanol | 5.10 | 5.03 (1) | 5.02 | 0.77 | 0.73 (5) | 0.73 | 89 | acetophenone | 4.93 | 5.11 (4) | 5.12 | 0.81 | 0.81 (1) | 0.81 |
| 15 | 1-decanol | 6.77 | 6.66 (2) | 6.64 | 0.70 | 0.69 (2) | 0.69 | 90 | nonanoic acid | 6.76 | 6.69 (1) | 6.67 | 0.52 | 0.52 (1) | 0.52 |
| 16 | 1-dodecanol | 8.26 | 8.06 (2) | 8.03 | 0.59 | 0.66 (11) | 0.66 | 91 | benzyl acetate | 5.73 | 6.27 (9) | 6.29 | 0.74 | 0.69 (7) | 0.69 |
| 17 | cycloheptylamine | 4.40 | 4.50 (2) | 4.51 | 0.79 | 0.81 (2) | 0.81 | 92 | 2-isopropylphenol | 6.11 | 6.05 (1) | 6.04 | 0.81 | 0.77 (5) | 0.77 |
| 18 | cyclohexylamine | 3.05 | 3.60 (18) | 3.63 | 0.78 | 0.81 (4) | 0.81 | 93 | benzyl cyanide | 5.45 | 5.13 (6) | 5.12 | 0.81 | 0.84 (3) | 0.84 |
| 19 | di-*n*-butylamine | 4.18 | 4.29 (3) | 4.30 | 0.75 | 0.79 (5) | 0.79 | 94 | phenylacetic acid | 6.54 | 6.50 (1) | 6.50 | 0.63 | 0.57 (10) | 0.56 |
| 20 | dodecylamine | 8.14 | 7.90 (3) | 7.88 | 0.69 | 0.74 (8) | 0.75 | 95 | 4-phenylbutyric acid | 7.86 | 8.20 (4) | 8.22 | 0.49 | 0.57 (16) | 0.58 |
| 21 | 1-bromodecane | 7.42 | 7.60 (2) | 7.62 | 0.60 | 0.52 (14) | 0.51 | 96 | thianisole | 5.16 | 4.58 (11) | 4.55 | 0.78 | 0.76 (2) | 0.76 |
| 22 | 1-methylpiperazine | 2.98 | 3.13 (5) | 3.11 | 0.57 | 0.56 (1) | 0.56 | 97 | benzyl methyl sulfide | 5.88 | 5.50 (6) | 5.47 | 0.74 | 0.77 (4) | 0.77 |
| 23 | 1,3-propanediol | 2.65 | 2.28 (14) | 2.21 | 0.45 | 0.45 (0) | 0.45 | 98 | 4-picoline | 2.98 | 2.96 (1) | 2.96 | 0.86 | 0.87 (1) | 0.87 |
| 24 | 1,4-butanediol | 3.72 | 2.64 (29) | 2.56 | 0.47 | 0.53 (13) | 0.54 | 99 | 3-picoline | 2.97 | 2.64 (11) | 2.63 | 0.86 | 0.87 (1) | 0.87 |
| 25 | chlorocyclohexane | 3.36 | 3.32 (1) | 3.40 | 0.71 | 0.68 (5) | 0.68 | 100 | *m*-cresol | 5.03 | 5.03 (0) | 5.03 | 0.77 | 0.77 (0) | 0.77 |
| 26 | cycloheptanone | 4.42 | 4.07 (8) | 4.05 | 0.78 | 0.77 (2) | 0.77 | 101 | hexamethylbenzene | 8.21 | 7.89 (4) | 7.86 | 1.04 | 1.02 (2) | 1.02 |
| 27 | cyclohexane carboxylic acid | 5.62 | 5.45 (3) | 5.44 | 0.60 | 0.53 (12) | 0.52 | 102 | 3-cyanopyridine | 4.19 | 4.09 (2) | 4.09 | 0.77 | 0.69 (10) | 0.69 |
| | | | | | | | | 103 | 2-isopropoxyphenol | 5.95 | 6.57 (10) | 6.60 | 0.72 | 0.66 (8) | 0.66 |
| 28 | naphthalene | 6.09 | 6.24 (2) | 6.24 | 1.07 | 0.96 (11) | 0.95 | 104 | 4-cyanopyridine | 3.89 | 3.83 (2) | 3.82 | 0.67 | 0.70 (5) | 0.70 |
| 29 | bibenzyl | 7.60 | 8.46 (11) | 8.53 | 0.90 | 0.91 (2) | 0.92 | 105 | 3-ethylpyridine | 3.92 | 3.65 (7) | 3.64 | 0.86 | 0.87 (1) | 0.87 |
| 30 | propylbenzene | 4.04 | 4.07 (1) | 4.07 | 1.05 | 1.06 (1) | 1.06 | 106 | 2,4,6-trimethylpyridine | 4.37 | 4.85 (11) | 4.86 | 0.87 | 0.88 (2) | 0.89 |
| 31 | allylbenzene | 3.93 | 4.06 (3) | 4.06 | 1.04 | 1.03 (1) | 1.03 | 107 | 2-amino-4,6-dimethyl-pyridine | 5.85 | 6.24 (7) | 6.24 | 0.66 | 0.70 (7) | 0.71 |
| 32 | indan | 4.79 | 5.12 (7) | 5.13 | 1.13 | 1.04 (8) | 1.04 | | | | | | | | |
| 33 | cumene | 3.75 | 4.03 (7) | 4.03 | 1.12 | 1.06 (5) | 1.06 | 108 | *trans*-cinnamaldehyde | 6.63 | 6.30 (5) | 6.29 | 0.72 | 0.78 (8) | 0.78 |
| 34 | phenathrene | 10.27 | 10.12 (1) | 10.13 | 0.79 | 0.83 (5) | 0.83 | 109 | dihydrocoumarin | 7.46 | 7.73 (4) | 7.75 | 0.60 | 0.67 (12) | 0.67 |
| 35 | tetralin | 5.92 | 6.00 (1) | 6.01 | 1.01 | 1.03 (2) | 1.03 | 110 | *trans*-cinnamic acid | 7.86 | 8.03 (2) | 8.04 | 0.50 | 0.54 (8) | 0.54 |
| 36 | 1,2,3,4-tetrahydro-quinoline | 7.17 | 6.75 (6) | 6.74 | 0.85 | 0.84 (1) | 0.84 | 111 | coumarin | 7.87 | 8.10 (3) | 8.12 | 0.65 | 0.63 (4) | 0.62 |
| | | | | | | | | 112 | 4-ethylpyridine | 3.97 | 3.88 (2) | 3.88 | 0.82 | 0.87 (6) | 0.87 |
| 37 | quinoline | 6.45 | 6.38 (1) | 6.38 | 0.82 | 0.80 (2) | 0.80 | 113 | 4-(2-aminoethyl) pyridine | 6.08 | 6.02 (1) | 6.01 | 0.72 | 0.73 (2) | 0.73 |
| 38 | mesitylene | 4.18 | 4.50 (8) | 4.50 | 1.15 | 1.07 (7) | 1.06 | 114 | ethyl pipecolinate | 6.11 | 7.25 (19) | 7.31 | 0.53 | 0.49 (8) | 0.48 |
| 39 | 1,2,3,4-tetrahydro-isoquinoline | 6.75 | 6.90 (2) | 6.91 | 0.85 | 0.85 (1) | 0.85 | 115 | 2-ethylpyridine | 3.44 | 3.39 (1) | 3.39 | 0.80 | 0.87 (8) | 0.87 |
| | | | | | | | | 116 | 2-(2-hydroxyethyl) pyridine | 5.63 | 5.72 (2) | 5.71 | 0.60 | 0.67 (12) | 0.67 |
| 40 | anthracene | 10.34 | 10.12 (2) | 10.10 | 0.76 | 0.83 (9) | 0.83 | 117 | benzil | 10.31 | 10.38 (1) | 10.40 | 0.59 | 0.56 (5) | 0.56 |
| 41 | α-methylstyrene | 4.27 | 4.22 (1) | 4.22 | 1.03 | 1.04 (0) | 1.04 | 118 | ethyl isonicotinate | 6.04 | 6.26 (4) | 6.27 | 0.58 | 0.59 (1) | 0.59 |
| 42 | dibenzofuran | 8.60 | 9.32 (8) | 9.36 | 0.66 | 0.69 (4) | 0.69 | 119 | 2-acetylpyridine | 4.60 | 4.98 (8) | 4.98 | 0.69 | 0.69 (0) | 0.69 |
| 43 | dibenzothiophene | 10.11 | 10.11 (0) | 10.11 | 0.71 | 0.64 (10) | 0.64 | 120 | 3-acetylpyridine | 5.20 | 4.61 (11) | 4.60 | 0.68 | 0.69 (1) | 0.69 |
| 44 | 1,1′-binaphthyl | 14.14 | 14.19 (0) | 14.26 | 0.76 | 0.68 (10) | 0.67 | 121 | isochroman | 6.04 | 6.01 (1) | 6.01 | 0.82 | 0.82 (0) | 0.82 |
| 45 | 1-tetralone | 7.48 | 7.52 (0) | 7.52 | 0.75 | 0.78 (4) | 0.78 | 122 | furfurylamine | 2.73 | 3.49 (28) | 3.53 | 0.58 | 0.62 (6) | 0.62 |
| 46 | 1-bromonaphthalene | 8.33 | 8.75 (5) | 8.79 | 0.47 | 0.51 (9) | 0.52 | 123 | 1-methyl-2-pyridone | 5.78 | 4.86 (16) | 4.82 | 0.48 | 0.52 (8) | 0.52 |
| 47 | 1-naphthol | 8.41 | 8.49 (1) | 8.48 | 0.58 | 0.67 (15) | 0.67 | 124 | phenanthridine | 10.49 | 10.22 (3) | 10.23 | 0.66 | 0.70 (6) | 0.70 |
| 48 | cyclohexylbenzene | 7.24 | 7.13 (2) | 7.13 | 0.96 | 1.00 (4) | 1.00 | 125 | 2-ethylphenol | 5.58 | 5.63 (1) | 5.63 | 0.82 | 0.77 (6) | 0.77 |
| 49 | 1-benzylnaphthalene | 11.39 | 11.46 (1) | 11.48 | 0.89 | 0.80 (11) | 0.79 | 126 | 3-pyridine carboxaldehyde | 4.17 | 4.73 (13) | 4.75 | 0.70 | 0.64 (8) | 0.64 |
| 50 | phenol | 4.16 | 3.99 (4) | 3.98 | 0.73 | 0.76 (4) | 0.76 | 127 | acridine | 10.35 | 10.12 (2) | 10.11 | 0.68 | 0.70 (3) | 0.70 |
| 51 | benzyl phenyl ether | 8.87 | 8.49 (4) | 8.46 | 0.62 | 0.74 (19) | 0.74 | 128 | diphenylamine | 9.16 | 9.21 (1) | 9.21 | 0.69 | 0.67 (3) | 0.67 |
| 52 | diphenyl ether | 7.74 | 7.69 (1) | 7.71 | 0.86 | 0.74 (14) | 0.73 | 129 | *p*-cresol | 5.02 | 4.68 (7) | 4.66 | 0.77 | 0.77 (0) | 0.77 |
| 53 | benzonitrile | 4.10 | 4.43 (8) | 4.44 | 0.91 | 0.83 (8) | 0.83 | 130 | 2-picoline | 2.56 | 2.52 (1) | 2.52 | 0.88 | 0.87 (1) | 0.87 |
| 54 | diphenyl sulfide | 9.02 | 8.99 (0) | 8.99 | 0.62 | 0.68 (9) | 0.68 | 131 | diphenylmethane | 7.96 | 7.75 (3) | 7.74 | 0.85 | 0.92 (8) | 0.92 |
| 55 | 5,6,7,8-tetrahydro-1-naphthol | 8.02 | 7.98 (1) | 7.98 | 0.82 | 0.73 (10) | 0.73 | 132 | bibenzyl | 8.61 | 8.57 (1) | 8.58 | 0.90 | 0.91 (1) | 0.91 |
| | | | | | | | | 133 | stilbene | 9.79 | 9.31 (5) | 9.30 | 0.76 | 0.85 (11) | 0.85 |
| 56 | cyclohexyl phenyl ketone | 9.14 | 9.33 (2) | 9.35 | 0.75 | 0.76 (1) | 0.76 | 134 | *p*-chlorotoluene | 4.03 | 4.26 (6) | 4.26 | 0.78 | 0.74 (5) | 0.74 |
| 57 | *n*-methylaniline | 4.94 | 4.52 (9) | 4.51 | 0.87 | 0.83 (5) | 0.83 | 135 | α-bromotoluene | 5.24 | 4.94 (6) | 4.91 | 0.56 | 0.59 (5) | 0.59 |
| 58 | 1,6-dihydroxynaphthalene | 10.33 | 9.62 (7) | 9.57 | 0.58 | 0.56 (4) | 0.56 | 136 | 2,4-dimethyl-3-pentanone | 2.43 | 2.94 (21) | 2.97 | 0.78 | 0.79 (1) | 0.79 |
| 59 | thiophene | 1.49 | 1.44 (4) | 1.41 | 0.67 | 0.68 (1) | 0.68 | 137 | benzene | 1.46 | 1.41 (3) | 1.41 | 1.09 | 1.10 (1) | 1.10 |
| 60 | pyrrole | 2.04 | 2.10 (3) | 2.09 | 0.74 | 0.67 (10) | 0.65 | 138 | toluene | 2.23 | 2.24 (0) | 2.24 | 1.17 | 1.10 (6) | 1.09 |
| 61 | benzylamine | 4.47 | 4.77 (7) | 4.77 | 0.85 | 0.87 (2) | 0.87 | 139 | *o*-xylene | 3.43 | 3.86 (12) | 3.88 | 1.04 | 1.08 (4) | 1.09 |
| 62 | *p*-nitrotoluene | 6.24 | 6.32 (1) | 6.33 | 0.66 | 0.61 (7) | 0.61 | 140 | decane | 4.59 | 4.46 (3) | 4.45 | 0.94 | 0.96 (2) | 0.96 |
| 63 | phenyl benzoate | 9.42 | 9.32 (1) | 9.33 | 0.69 | 0.60 (13) | 0.59 | 141 | eicosane | 11.55 | 11.64 (1) | 11.67 | 0.67 | 0.72 (8) | 0.73 |
| 64 | *p*-toluidine | 5.02 | 4.91 (2) | 4.91 | 0.86 | 0.84 (3) | 0.84 | 142 | cyclohexane | 1.53 | 0.92 (40) | 0.83 | 1.04 | 1.04 (0) | 1.04 |
| 65 | fluorene | 10.28 | 9.22 (10) | 9.17 | 0.86 | 0.87 (1) | 0.87 | 143 | heptane | 1.78 | 1.45 (19) | 1.42 | 1.00 | 1.03 (3) | 1.04 |
| 66 | 4-ethylphenol | 5.81 | 5.54 (5) | 5.53 | 0.78 | 0.77 (2) | 0.77 | 144 | ethylbenzene | 3.12 | 3.26 (5) | 3.27 | 0.93 | 1.08 (16) | 1.09 |
| 67 | aniline | 4.08 | 4.19 (3) | 4.18 | 0.82 | 0.82 (0) | 0.82 | 145 | propiophenone | 5.82 | 5.54 (5) | 5.53 | 0.84 | 0.83 (1) | 0.83 |
| 68 | *o*-toluidine | 5.00 | 5.30 (6) | 5.30 | 0.86 | 0.84 (3) | 0.84 | 146 | anisole | 3.59 | 3.39 (6) | 3.39 | 0.79 | 0.84 (6) | 0.84 |
| 69 | butyl phenyl ether | 6.08 | 5.79 (5) | 5.78 | 0.84 | 0.81 (3) | 0.81 | 147 | 1,7-dihydroxynaphthalene | 10.30 | 9.71 (6) | 9.67 | 0.50 | 0.57 (13) | 0.57 |
| 70 | bromobenzene | 3.75 | 4.08 (9) | 4.09 | 0.55 | 0.56 (2) | 0.56 | 148 | isoquinoline | 6.65 | 6.51 (2) | 6.51 | 0.82 | 0.80 (3) | 0.80 |
| 71 | iodobenzene | 4.74 | 4.46 (6) | 4.44 | 0.43 | 0.46 (7) | 0.47 | 149 | 2,2′-bipyridine | 7.98 | 8.17 (2) | 8.19 | 0.73 | 0.69 (5) | 0.69 |
| 72 | benzaldehyde | 3.92 | 4.42 (13) | 4.43 | 0.81 | 0.80 (1) | 0.80 | 150 | 2,4,6-trimethylbenzoic acid | 7.82 | 8.30 (6) | 8.32 | 0.57 | 0.60 (4) | 0.60 |
| 73 | benzyl alcohol | 4.64 | 4.93 (6) | 4.94 | 0.86 | 0.79 (8) | 0.79 | 151 | methyl phenyl sulfoxide | 7.07 | 5.98 (15) | 5.87 | 0.58 | 0.60 (3) | 0.60 |
| 74 | *p*-tolualdehyde | 5.10 | 5.27 (3) | 5.27 | 0.81 | 0.81 (0) | 0.81 | 152 | 4-phenyl-1-butene | 4.87 | 4.41 (9) | 4.38 | 1.08 | 1.05 (3) | 1.05 |
| 75 | 1,4-naphthoquinone | 7.68 | 8.35 (9) | 8.41 | 0.68 | 0.61 (10) | 0.61 | | | | | | | | |

and their errors $\Delta c_i$ are determined by the least-squares method. The model described by eq 1 is obtained by fitting experimental values of a property $P$ to linear (or parabolic, or cubic) functions of initial descriptors. Parabolic or cubic or higher order curves are tested provided the second-order, third-order, and higher order terms of the initial descriptors are included. The cross-product terms correspond to the synergistic relation between descriptors (topological, geometrical, quantum-chemical, or structural properties of molecules). Using eq 1, one can estimate the values $P^{est}$ for property $P$ for all of the molecules. The $P^{est}$ and $P$ values are not the same, and their comparison defines the standard error of estimate $S$ and the coefficient of multivariate correlation $R$, which measures the quality of the fit. The quality of prediction for a new molecule from the same class is indicated by the cross-validated correlation coefficient ($R_{cv}$) and the cross-validated standard error of estimate ($S_{cv}$). If too many descriptors are introduced into the MR model, the statistical parameters $R$ and $S$ will increase and decrease, respectively, while the cross-validated parameters $R_{cv}$ and $S_{cv}$ will decrease and increase. This demonstrates that a good MR model must be simple to guarantee the reliable prediction for a new molecule of the property studied. In practice, it is necessary to select $I$ ($I = 1, ..., K$, usually $K \leq 10$) significant descriptors in MR models from the initial set of $N$ descriptors, where $N$ can be very large (in this paper $N = 296$).

It is important to note that all of the values of $S$ and $S_{cv}$ given in this paper were estimated with factor $M$ as the denominator. For consistency and for a direct comparison to be made, all $S$ values from ref 8 given in this paper are now expressed in the same manner (in ref 8 factor $M - I - 1$ was used as the denominator for estimating $S$, where $M$ is the number of examples (molecules) and $I$ denotes the number of descriptors used in MR model).

**Selection of the Best Possible Descriptors in MR Models.** Once the large set of descriptors has been calculated, the main problem is the selection of that restricted set which contains the most informative descriptors for modeling the property. Following Okham's Razor,[21] the search for the best model (in terms of maximum $R$ and minimum $S$ values) should start from the simplest and proceed to more complex multiregression models.

To be able to do that, we developed a computer program for the selection of the best possible descriptors in MR models. The number of descriptors that can be selected in a such way in a MR model is limited by the size of the descriptor data set. In the case of the data set with 296 descriptors, we were be able to select the best possible MR models with four (in several hours on HP 9000/E55) or even with five (in approximately 4 days on SUN Enterprise 3000) descriptors. For example, there are $3 \times 10^8$ possible models with four descriptors that one can select out of 296 descriptors (296!/(296 − 4)!4!).

The quality of each model (with $I$ descriptors) was identified with its correlation coefficient ($R$), and among all possible models the best one (with the highest value of $R$) was selected. To be able to check the quality of a large number of MR models, it was necessary to develop a very fast procedure for calculating $R$, which was achieved by the orthogonalization of descriptors. Namely, in the case in which one has the MR model based on the set of $I$ orthogonalized descriptors $d_i$ ($i = 1, ..., I$), the correlation coefficient between the experimental values of modeled property $P$ and the values estimated by the model $P^{est}$ can be calculated in a very simple way:

$$R = [\sum_{i=1}^{I} R_i^2]^{1/2} \tag{2}$$

where $R_i$ is the correlation coefficient between each orthogonalized descriptor $d_i$ and the modeled property $P$.

**Stepwise Selection of Descriptors.** The procedure for selecting the best possible model with five descriptors out of 296 descriptors could not be used as a starting model to obtain the best possible model with six or more descriptors. For this purpose, a new procedure for the stepwise selection of descriptors in the MR model was developed. Most (or even all) procedures for stepwise variable/descriptor selection have hitherto been based on the introduction of novel descriptors in each step by adding additional descriptors *one* by *one* to those already selected (*1-way* or *1-fold* stepwise variable selection). Starting from the optimum MR model with $I$ descriptors (we tried to reach as many as possible values of $I$), in each subsequent step a number $i$ of the best additional descriptors was combined with those previously selected. We name this stepwise variable selection $i$ by $i$, or as an $i$-fold stepwise variable selection. Importantly, these $i$ newly selected descriptors, with all the previously selected descriptors, give the MR model which is the best of all those that can be obtained by addition of any other $i$-tuple of descriptors to the previously selected descriptors.

**Orthogonalization of Descriptors.** The orthogonalization of $I$ descriptors selected in the MR model was carried out by the application of eq 3 which describes the orthogonalization of vector $|k\rangle$ against vector $|j\rangle$,[34] where $|k'\rangle$ is the descriptor $|k\rangle$ orthogonalized against descriptor $|j\rangle$, while $r_{jk}$ is the correlation coefficient between the nonorthogonal descriptors $|j\rangle$ and $|k\rangle$. After this transformation descriptor $|j\rangle$ remains unchanged, as does the correlation coefficient between $|j\rangle$ and property $|P\rangle$, since $|j\rangle$ is the first descriptor against which the second descriptor was orthogonalized. The same procedure was repeated *for all the remaining descriptors* in the order in which we want them to be orthogonalized against the first and, then, against the second, third, ..., and ($I - 1$) descriptor.

$$|k'\rangle = \frac{r_{jk}|j\rangle - |k\rangle}{[1 - r_{jk}^2]^{1/2}} \tag{3}$$

The correlation coefficient between property $|P\rangle$ and the orthogonalized descriptor $|k'\rangle$ is now changed. The new correlation coefficient $R'_2$ can be computed by

$$R'_2 = \langle P|k'\rangle = \frac{r_{jk}\langle P|j\rangle - \langle P|k\rangle}{[1 - r_{jk}^2]^{1/2}} = \frac{r_{jk}R_j - R_k}{[1 - r_{jk}^2]^{1/2}} \tag{4}$$

where $\langle P|$ is a bra vector representation of property $P$. $R_j$, $R_k$, and $r_{jk}$ stand, respectively, for the correlation coefficients between property $\langle P|$ and nonorthogonal descriptors $|j\rangle$ and $|k\rangle$, and that between the nonorthogonal descriptors $|j\rangle$ and $|k\rangle$. The orthogonalization of the descriptors by subsequent application of eq 3 was continued until each of them is orthogonal to all others. All the computations regarding eqs

3 and 4 were achieved solely by considering the initially calculated correlation coefficients between the nonorthogonalized descriptors. It is important to note that the descriptors in the final MR model are nonorthogonalized.

**Reducing the Number of Descriptor in the Data Set.** For modeling the GC retention times and response factors, it was not possible to obtain the best MR models with six or seven descriptors from the whole set of 296 descriptors. Even obtaining the best possible MR model with five descriptors needed about 4 days CPU time on a SUN Enterprise 3000 (UltraSPARC processor, 250 MHz). However, the best possible MR model with 4 out of 296 descriptors was selected ~50 times (~2 h) faster than the model with five descriptors. Therefore, the following strategy was applied for our variable selection of descriptors: (1) determination of the optimum MR models with one, two, three, and four descriptors; (2) determination of the MR models with 25−30 descriptors by application of the *i*-fold stepwise procedure (for $i = 1, 2, 3, 4$), starting from each of the models in 1 as the initial one; (3) after step 2 was completed, 16 MR models (four initial models by four step-values) with 25−30 descriptors were obtained. Because several descriptors were shared between many MR models, a novel reduced data set of descriptors was obtained in such a way that each descriptor was taken into account only once.

The reduced data sets for RT and RF modeling contained 100 and 96 descriptors, respectively. For such data sets it was possible to select the best MR models with five, six, and seven descriptors.

**Introduction of Nonlinear Terms of Descriptors.** After the elucidation of the best MR models with 1, 2, ..., 7 descriptors, it was logical to search for nonlinear dependencies between the modeled properties and the descriptors that showed up in these models. The simplest way to introduce nonlinearities is through cross-product terms of the initial descriptors. This is also the most convenient way, since each nonlinear function (if we assume that it is an analytic function) can be expressed as a Taylor series of the initial descriptors. Moreover, very often the most important terms are of the linear and parabolic forms. Other modeling techniques that are used in chemistry also take into account nonlinearities; this is especially true in the case of neural networks (see for example refs 24 and 35).

In the present work, the cross-product terms of all descriptors that enter the best MR models with 1, 2, ..., 7 descriptors were calculated. Then, all the initial (linear) descriptors were considered together with all the nonlinear descriptors. Thus, by introducing the cross-products of data set with *k* descriptors, a new data set which contained *k* linear and $(k + 1)(k/2)$ nonlinear descriptors was obtained. The selection of descriptors for the best (nonlinear) MR models from this data set was carried out by applying the same procedure as that described in the above case on the 296-descriptor set.

## RESULTS AND DISCUSSION

To obtain the MR models for gas chromatographic retention times ($t_R$) and response factors, two classes of linear models were obtained from the initial data set of 296 descriptors: (1) the optimum MR models containing from one to five descriptors and (2) the best stepwise MR models with six and seven descriptors that were obtained after preselection of descriptors. These models are directly comparable to corresponding MR models with the same number of descriptors, obtained by the CODESSA program.[8] Statistical parameters of the best two- to seven-descriptor models for $t_R$ and RF, obtained by the CODESSA program, are given in Table 2a,b, respectively. Details of the best six-decriptor models are also shown in Table 2.

**Table 2.** Best Multiregression Models with Two to Six Descriptors for Gas Chromatographic Retention Times ($t_R$) and Response Factors (RF) of 152 Compounds Obtained by the Statistical Treatment Utilizing the CODESSA Program

(a) Retention Times

| I[a] | $R^2$ [b] | $R_{cv}^2$ [c] | S[d] |
|---|---|---|---|
| 2 | 0.9134 | 0.9092 | 0.7313 |
| 3 | 0.9318 | 0.9271 | 0.6489 |
| 4 | 0.9465 | 0.9419 | 0.5744 |
| 5 | 0.9539 | 0.9496 | 0.5336 |
| 6 | 0.9590 | 0.9550 | 0.5032 |

Best Six-Descriptor Model of $t_R$

| | |
|---|---|
| intercept | (26.50 ± 5.98) |
| relative no. of C−H bonds | (−6.912 ± 0.746) |
| total entrophy at 300 K/no. of atoms | (−0.871 ± 0.102) |
| α polarizability | (0.046 ± 0.005) |
| molecular weight | (0.018± 0.003) |
| min valency of a H atom | (−21.55 ± 4.02) |
| max atomic orbital electronic population | (0.929 ± 0.218) |

(b) Response Factors

| I[a] | $R^2$ [b] | $R_{cv}^2$ [c] | S[d] |
|---|---|---|---|
| 2 | 0.7630 | 0.7531 | 0.0787 |
| 3 | 0.8256 | 0.8160 | 0.0675 |
| 4 | 0.8490 | 0.8388 | 0.0628 |
| 5 | 0.8869 | 0.8763 | 0.0544 |
| 6 | 0.8924 | 0.8810 | 0.0530 |

Best Six-Descriptor Model of the RF

| | |
|---|---|
| intercept | (−2.327 ± 0.458) |
| relative weight of "effective" C atoms | (−0.958 ± 0.046) |
| total mol one-center one-electron repulsion energy | (−0.0028 ± 0.0002) |
| relative no. of "effective" C atoms | (−1.160 ± 0.102) |
| min total bond order (>0.1) of a C atom | (−0.206 ± 0.019) |
| min valency of a H atom | (3.316 ± 0.376) |
| total hybridization component of the molecular dipole | (−0.030 ± 0.011) |

[a] Number of descriptors in the multiregression models. [b] Square of the correlation coefficient. [c] Square of the cross-validated (leave-one-out) correlation coefficient. [d] Standard error of estimate (root mean square error).

Starting from descriptors that were selected in all linear models (up to the model with seven descriptors) and taking into account their cross-product terms, we then obtained the optimum and the best stepwise nonlinear MR models for $t_R$ and RF. Descriptors involved in all of the MR models are listed in Table 3.

**Multiregression Models of Retention Times. (a) Best Linear $t_R$ Models.** The best linear MR models with one up to five descriptors were selected from the data set containing 296 descriptors and are given in Table 4. In the first step, the best possible models for $t_R$ with $I = 1, 2, 3$, and 4 descriptors were selected from the whole set of 296 descriptors.

The new (*i* by *i*) stepwise MR procedure was also applied for the *preselection* of the most important descriptors out of

**Table 3.** Molecular Descriptors Involved in the Multiregression Models for Gas Chromatographic Retention Times ($t_R$) and Response Factors (RF) Selected by the Application of Novel Procedures for Descriptors Selection

| no. [a] | descriptors |
|---|---|
| | (a) Descriptors Selected in Modeling of Retention Times |
| 6 | no. of C−H bonds |
| 8 | no. of C−X bonds |
| 9 | rel no. of C−H bonds |
| 27 | av atomic nucleophilic reaction index for a C atom |
| 51 | total dipole of the molecule |
| 52 | Image of the Onsager−Kirkwood solvation energy |
| 66 | min valency of a H atom |
| 101 | max exchange energy for a C−H bond |
| 137 | highest normal mode vibrational freq |
| 150 | rotational entropy of the molecule at 300 K |
| 206 | gravitation index (all bonds) |
| 212 | Randic index (order 3) |
| 216 | Kier and Hall index (order 3) |
| 248 | XY shadow |
| 252 | ZX shadow |
| | (b) Descriptors Selected in Modeling of Response Factors |
| 5 | rel weight of C atoms |
| 9 | rel no. of C−H bonds |
| 13 | rel no. of "effective" C atoms |
| 15 | rel weight of "effective" C atoms |
| 54 | max atomic orbital electronic population |
| 62 | av valency of a C atom |
| 63 | min total bond order (>0.1) of a C atom |
| 65 | av total bond order of a C atom |
| 66 | min valency of a H atom |
| 82 | min atomic state energy for a H atom |
| 114 | total molecular one-center electron−electron repulsion |
| 128 | ALFA polarizability (DIP) |
| 210 | Randic index (order 1) |
| 248 | XY shadow |
| 263 | polarity parameter ($Q_{max} - Q_{min}$) |
| 281 | fractional negative charge weighted surface area (FNSA-2 = total charge weighted negative surface area/(total molecular surface area)) |

[a] Numbering of descriptors corresponds to that in 296-descriptor data set.

296. To the best possible models with 1, 2, 3, and 4 descriptors, the new *i* descriptors were added, and new MR models were obtained in each next step. This was repeated up to the MR models with 25−30 descriptors. In this case the number of new-added descriptors was varied from $i = 1$ to $i = 4$. As a final result 16 MR models with 25−30 descriptors were produced. The descriptors from these models gave a novel data set. In the case of $t_R$, the data set with 100 descriptors was obtained (several descriptors were involved in more than one model). The new data set of preselected descriptors for $t_R$ was used for selecting the best possible MR models with 5, 6, and 7 descriptors, that are given in Table 4.

To verify the efficiency of the *i* by *i* stepwise selection procedure, the selection of the optimum MR models with five descriptors (out of 296) was undertaken. The computation on the SUN Enterprise 3000 computer was completed after 4 days, and the model obtained was exactly the same as the model obtained after preselection of descriptors by application of the *i* by *i* stepwise procedure. This result is only related to this experiment, and, generally, it will be different.

Significantly, each of the models with $I = 1, ..., 6$ descriptors shares one or more descriptors with the other models (see Table 4). In the model with seven descriptors five novel descriptors (that do not appear in any other model) are involved but no descriptor from the model with six descriptors. However, as can be seen from Tables 3 and 4, some of the five novel descriptors are similar to those involved in other models (with $I \leq 6$). Notably, very good models are obtained containing one and two descriptors. Both of these models involve the *gravitation index* calculated over all bonds according to ref 31. This index is a measure of the compactness of the molecular 3D mass distribution. In other words, the *gravitation index* encodes simultaneously information about both the mass and the shape of the molecule, which are the most important properties for estimation of the retention time of a molecule.

Comparison of Tables 2 and 4 demonstrates that models with the same number of descriptors, obtained by the procedure described in this paper, are consistently better. Furthermore, the seven-descriptor model from Table 4 is significantly better than the six-descriptor models from Tables 2 and 4. It is very important to note that the seven-descriptor model contains five descriptors not involved in the models with $I \leq 6$. The descriptors from this seven-descriptor model, as well as those from the models with $I \leq 6$, are all utilized in the search for improved nonlinear MR models.

**(b) Best Nonlinear MR Models of $t_R$.** In the best linear MR models given in Table 4 (with $I \leq 7$) 15 descriptors

**Table 4.** Best Possible One- to Five-Descriptor and "the Best" Stepwise Six- and Seven-Descriptor Linear Multiregression Models for Gas Chromatographic Retention Times ($t_R$) of 152 Compounds Selected from 296 Descriptors

| $I$[a] | $D$[b] | $R^2$[c] | $R_{cv}^2$[d] | $S$[e] | $S_{cv}$[f] |
|---|---|---|---|---|---|
| 1 | *d*206 | 0.9007 | 0.8980 | 0.783 | 0.793 |
| 2 | *d*137, *d*206 | 0.9347 | 0.9318 | 0.635 | 0.649 |
| 3 | *d*66, *d*150, *d*212 | 0.9437 | 0.9406 | 0.589 | 0.605 |
| 4 | *d*66, *d*101, *d*150, *d*212, | 0.9541 | 0.9510 | 0.532 | 0.550 |
| 5 | *d*9, *d*27, *d*66, *d*212, *d*248 | 0.9644 | 0.9612 | 0.469 | 0.490 |
| 6 | *d*9, *d*27, *d*51, *d*66, *d*212, *d*248 | 0.9676 | 0.9641 | 0.447 | 0.471 |
| | $t_R = (26.9 \pm 3.1) + (-5.07 \pm 0.72)d9 + (-121.6 \pm 13.5)d27 + (0.14 \pm 0.037)d51 + (-26.0 \pm 3.4)d66 + (0.97 \pm 0.070)d212 + (0.094 \pm 0.0053)d248$ | | | | |
| 7 | *d*6, *d*8, *d*52, *d*137, *d*150, *d*216, *d*252 | 0.9732 | 0.9697 | 0.407 | 0.433 |
| | $t_R = (-30.7 \pm 1.5) + (-0.54 \pm 0.022)d6 + (0.44 \pm 0.028)d8 + (8.45 \pm 0.95)d52 + (0.0036 \pm 0.00027)d137 + (0.74 \pm 0.057)d150 + (0.85 \pm 0.10)d216 + (-0.056 \pm 0.0094)d252$ | | | | |

[a] Number of descriptors in the MR model. [b] Descriptors involved in the MR model; descriptor numbering corresponds to that in Table 3. [c] Square of the correlation coefficient. [d] Square of the leave-one-out cross-validated correlation coefficient. [e] Standard error of estimate. [f] Leave-one-out cross-validated standard error of estimate.

**Table 5.** Details of the Optimum One- to Six-Descriptor and the Best Stepwise Seven-Descriptor Nonlinear Multiregression Models for Gas Chromatographic Retention Times ($t_R$) of 152 Compounds Selected from 135 Descriptors (15 Descriptors Selected in Models Given in Table 3 and Their Cross-Products)

| $I$[a] | $D$[b] | $R^2$[c] | $R_{cv}^2$[d] | $S$[e] | $S_{cv}$[f] |
|---|---|---|---|---|---|
| 1 | *d*137·*d*206 | 0.9320 | 0.9300 | 0.6476 | 0.6572 |
| 2 | *d*137·*d*206, *d*206·*d*252 | 0.9418 | 0.9389 | 0.5996 | 0.6139 |
| 3 | *d*66·*d*101, *d*66·*d*150, *d*66·*d*212 | 0.9543 | 0.9517 | 0.5313 | 0.5461 |
| 4 | *d*212, *d*248, *d*6·*d*27, *d*66·*d*101 | 0.9624 | 0.9596 | 0.4815 | 0.4993 |
| 5 | *d*8, *d*6·*d*137, *d*9·*d*52, *d*101·*d*216, *d*137·*d*150 | 0.9701 | 0.9666 | 0.4287 | 0.4541 |
| 6 | *d*6·*d*137, *d*8·*d*150, *d*8·*d*252, *d*9·*d*52, *d*137·*d*150, *d*137·*d*216 | 0.9767 | 0.9738 | 0.3790 | 0.4020 |
| | $t_R = (-14.14 \pm 0.62) + (-16.31 \pm 0.44) \times 10^{-5}\, d6{\cdot}d137 + (15.96 \pm 0.83) \times 10^{-3}\, d8{\cdot}d150 + (-24.0 \pm 3.1) \times 10^{-4}\, d8{\cdot}d252 + (18.8 \pm 2.0) d9{\cdot}d52 + (16.55 \pm 0.69) \times 10^{-5}\, d137{\cdot}d150 + (26.4 \pm 2.9) \times 10^{-5}\, d137{\cdot}d216$ | | | | |
| 7 | *d*8, *d*6·*d*137, *d*9·*d*52, *d*9·*d*216, *d*27·*d*248, *d*27·*d*252, *d*137·*d*150 | 0.9798 | 0.9771 | 0.3520 | 0.3763 |
| | $t_R = (-17.39 \pm 0.62) + (0.462 \pm 0.016) d8 + (-20.59 \pm 0.46) \times 10^{-5}\, d6{\cdot}d137 + (19.0 \pm 1.9) d9{\cdot}d52 + (1.63 \pm 0.20) d9{\cdot}d216 + (4.16 \pm 0.55) d27{\cdot}d248 + (-6.90 \pm 0.82) d27{\cdot}d252 + (20.02 \pm 0.70) \times 10^{-5}\, d137{\cdot}d150$ | | | | |

[a] Number of descriptors in the MR model. [b] Descriptors involved in the MR model; descriptor numbering corresponds to that in Table 3. [c] Square of the correlation coefficient. [d] Square of the leave-one-out cross-validated correlation coefficient. [e] Standard error of estimate. [f] Leave-one-out cross-validated standard error of estimate.

were involved. These 15 descriptors together with their cross-products comprise our nonlinear set of 135 descriptors. From these 135 descriptors the best possible (nonlinear) MR models with $I \leq 6$ descriptors were selected (Table 5).

Obtaining the best nonlinear MR model with seven descriptors required use of the stepwise (*i* by *i*) procedure for the descriptor selection. For this, the values of *i* were varied between 1 and 4, and the stepwise selection was started each time from *I* descriptors (models with $I = 1$, ..., 6 in Table 5) selected in the best possible nonlinear MR models. The stepwise *i* by *i* procedure was stopped when models containing 15−18 descriptors were obtained. Combination of step *i* ($i = 1$, ..., 4) with starting models containing *I* descriptors ($I = 1$, ..., 6) gave 24 nonlinear MR models with 15−18 descriptors. Several of the descriptors involved were shared between many nonlinear MR models so that a new reduced data set of 100 descriptors was obtained. The best possible MR model with seven descriptors (given in Table 5) was selected from this set by applying the procedure for selecting the best possible MR model. According to the calculated statistical parameters, this model is the best of all the MR models for retention times obtained in this paper. The best seven-descriptor $t_R$ model is tested in order to exclude the possibility of chance correlation. The dependent variable is randomly scrambled 1000 times, and the squares of the correlation coefficients of such models were between 0.010 and 0.095.

A plot of observed vs calculated (1-fold cross-validated) retention times using the seven-parameter model from Table 5 is shown in Figure 1.

One can see that the experimental and predicted values agree reasonably well. Of course, the accuracy of predicted values is limited by the model error. Although the calculated values were obtained through 1-fold cross-validation, comparison with Figure 1 in ref 8 clearly shows that the seven-descriptor model obtained in this paper is superior. Interestingly, in both the linear and the nonlinear MR models for retention times, the simple path three connectivity indices[36,37] (which measure the complexity of the bonding within the molecule) were frequently involved. Similar results were obtained in several models for retention times developed by the Jurs group for diverse sets of compounds.[3,5−7]

**Multiregression Models of Response Factors. (a) Best Linear RF Models.** The best linear MR models of response factors containing 1, 2, 3, 4, and 5 descriptors were selected from the data set containing 296 descriptors (Table 6). First, the best possible models for RF with $I = 1$, 2, 3, and 4 descriptors were selected from the whole set of 296 descriptors.

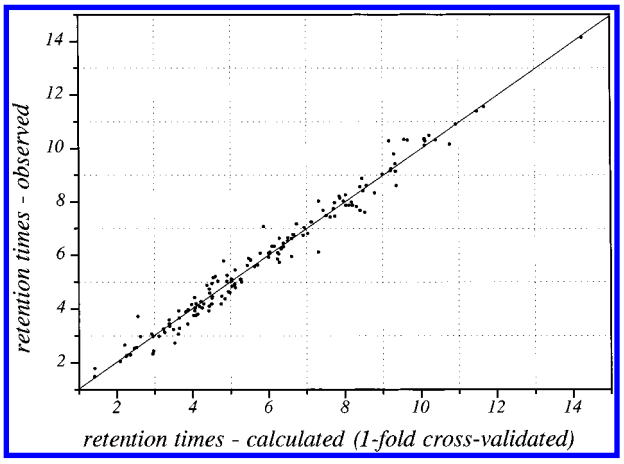


**Figure 1.** Plot of observed vs calculated retention times $t_R$ obtained using the best nonlinear seven-parameter multiregression model (model with $I = 7$ in Table 5).

In the second step the *preselection* of descriptors from the initial set was performed. For that, the *i* by *i* stepwise MR procedure was used. Starting from the optimum models with 1, 2, 3, and 4 descriptors, a set of *i* new descriptors was added in each next step. Together with the old descriptors these new-added descriptors produce new models, and the best according to the correlation coefficient was selected. This was repeated up to the MR models with 25−30 descriptors. The number of new descriptors that were added in each step was varied from $I = 1$ to $I = 4$. As a final result 16 MR models with 25−30 descriptors were produced. The descriptors used in these models were employed to create a new data set containing both linear and nonlinear (cross-products) descriptors. A data set with 96 descriptors was thus obtained. The new subset of preselected descriptors for RF was used to select the optimum MR models with 5, 6, and 7 descriptors. The statistical parameters of these models, and the names of descriptors that appear in them, are given in Tables 3 and 6, respectively.

**Table 6.** Best Possible One- to Five-Descriptor and "the Best" Stepwise Six- and Seven-Descriptor Linear Multiregression Models for Gas Chromatographic Response Factors (RF) of 152 Compounds Selected from 296 Descriptors

| $I^a$ | $D^b$ | $R^{2\,c}$ | $R_{cv}^{2\,d}$ | $S^e$ | $S_{cv}^f$ |
|---|---|---|---|---|---|
| 1 | *d*5 | 0.5145 | 0.5008 | 0.1127 | 0.1142 |
| 2 | *d*5, *d*114 | 0.7631 | 0.7531 | 0.0787 | 0.0803 |
| 3 | *d*5, *d*82, *d*210 | 0.8257 | 0.8147 | 0.0675 | 0.0696 |
| 4 | *d*5, *d*65, *d*128, *d*263 | 0.8626 | 0.8513 | 0.0599 | 0.0623 |
| 5 | *d*13, *d*15, *d*63, *d*66, *d*114 | 0.8870 | 0.8764 | 0.0544 | 0.0569 |
| 6 | *d*13, *d*15, *d*63, *d*66, *d*248, *d*281 | 0.8931 | 0.8816 | 0.0529 | 0.0556 |
| | $RF = (-2.19 \pm 0.42) + (-0.992 \pm 0.117)d13 + (0.986 \pm 0.054)d15 + (-0.223 \pm 0.019)d63 + (3.13 \pm 0.43)d66 + (-4.22 \pm 0.42) \times 10^{-3}\, d248 + (1.13 \pm 0.22)d281$ | | | | |
| 7 | *d*5, *d*9, *d*54, *d*62, *d*63, *d*66, *d*128 | 0.9022 | 0.8900 | 0.0506 | 0.0536 |
| | $RF = (-12.4 \pm 2.0) + (0.816 \pm 0.064)d5 + (0.383 \pm 0.081)d9 + (-0.118 \pm 0.024)d54 + (2.62 \pm 0.50) \times 10^{-3}\, d62 + (-0.153 \pm 0.020)d63 + (2.59 \pm 0.39)d66 + (-2.95 \pm 0.19) \times 10^{-3}\, d128$ | | | | |

[a] Number of descriptors in the MR model. [b] Descriptors involved in the MR model; descriptor numbering corresponds to that in Table 3. [c] Square of the correlation coefficient. [d] Square of the leave-one-out cross-validated correlation coefficient. [e] Standard error of estimate. [f] Leave-one-out cross-validated standard error of estimate.

**Table 7.** Details of the Best Possible One- to Six-Descriptor and "the Best" Stepwise Seven-Descriptor Nonlinear Multiregression Models for Gas Chromatographic Response Factors (RF) of 152 Compounds Selected from 152 Descriptors (16 Descriptors Selected in Models Given in Table 6 and Their Cross-Products)

| $I^a$ | $D^b$ | $R^{2\,c}$ | $R_{cv}^{2\,d}$ | $S^e$ | $S_{cv}^f$ |
|---|---|---|---|---|---|
| 1 | *d*5·*d*82 | 0.5780 | 0.5656 | 0.1050 | 0.1066 |
| 2 | *d*5·*d*82, *d*65·*d*114 | 0.8436 | 0.8370 | 0.0639 | 0.0653 |
| 3 | *d*5·*d*82, *d*65·*d*210, *d*65·*d*263 | 0.8656 | 0.8575 | 0.0593 | 0.0610 |
| 4 | *d*5·*d*15, *d*13·*d*248, *d*54·*d*63, *d*62·*d*66 | 0.8878 | 0.8789 | 0.0542 | 0.0563 |
| 5 | *d*13·*d*248, *d*15·*d*54, *d*54·*d*82, *d*62·*d*66, *d*63·*d*82 | 0.8981 | 0.8884 | 0.0516 | 0.0540 |
| 6 | *d*5·*d*62, *d*9·*d*15, *d*13·*d*248, *d*54·*d*210, *d*62·*d*66, *d*63·*d*65 | 0.9093 | 0.8995 | 0.0487 | 0.0513 |
| | $RF = (-2.39 \pm 0.35) + (0.175 \pm 0.017)d5{\cdot}d62 + (0.548 \pm 0.064)d9{\cdot}d15 + (-6.82 \pm 0.96) \times 10^{-3}\, d13{\cdot}d248 + (-17.8 \pm 2.5) \times 10^{-3}\, d54{\cdot}d210 + (0.726 \pm 0.090)d62{\cdot}d66 + (-0.160 \pm 0.016)d63{\cdot}d65$ | | | | |
| 7 | *d*5·*d*82, *d*9·*d*15, *d*13·*d*248, *d*54·*d*210, *d*62·*d*66, *d*63·*d*65, *d*82·*d*82 | 0.9146 | 0.9041 | 0.0472 | 0.0501 |
| | $RF = (-5.6 \pm 1.2) + (92.0 \pm 9.0) \times 10^{-3}\, d5{\cdot}d82 + (532.6 \pm 62.4) \times 10^{-3}\, d9{\cdot}d15 + (-6.95 \pm 0.93) \times 10^{-3}\, d13{\cdot}d248 + (-17.9 \pm 2.5) \times 10^{-3}\, d54{\cdot}d210 + (1.90 \pm 0.39)d62{\cdot}d66 + (-149.1 \pm 16.5) \times 10^{-3}\, d63{\cdot}d65 + (-24.3 \pm 6.3) \times 10^{-3}\, d82{\cdot}d82$ | | | | |

[a] Number of descriptors in the MR model. [b] Descriptors involved in the MR model; descriptor numbering corresponds to that in Table 3. [c] Square of the correlation coefficient. [d] Square of the leave-one-out cross-validated correlation coefficient. [e] Standard error of estimate. [f] Leave-one-out cross-validated standard error of estimate.

To verify the quality of the *i* by *i* stepwise selection procedure, the optimum MR model with 5 out of the 296 descriptors was determined after 4 days of computation on the SUN Enterprise 3000 computer. This model was exactly the same as that obtained by preselection of descriptors and application of the *i* by *i* stepwise procedure.

While several common descriptors appear in the models with $I \leq 6$ descriptors, three descriptors (*d*9, *d*54, *d*62) are involved only in the model with $I = 7$. Moreover, the models with four and five descriptors have no common descriptor, and the models with six and seven descriptors contain only two common descriptors (*d*63, minimum total bond order ($> 0.1$) of a C atom, and d66, minimum valency of a H atom). Altogether, 16 descriptors are involved in the seven MR models for RF. Among them, the most often used is *d*5 (relative weight of C atoms), as well as descriptors that express the relative weight or number of C atoms or C−H bonds, and those related to the valency or bond order of a C or H atom.

From the comparison between the models having the same number of descriptors listed in Tables 2 and 6, one can see that many MR models are of the same quality. However, the seven-descriptor model given in Table 6 is the best of all the models obtained. Furthermore, the comparison between the six-descriptor models from Tables 2 and 6 with the seven-descriptor model from Table 7 clearly shows that this seven-descriptor model is significantly better. It is important to note that this model contains three descriptors that are not involved in the models with $I \leq 6$. These new descriptors are also utilized in searching for the nonlinear MR models.

**(b) Best Nonlinear MR Models of RF.** With 16 descriptors involved in the linear MR models for RF (Table 6), as well as with their cross-products, a mixed linear−nonlinear set with 152 descriptors was obtained. From this set the optimum (nonlinear) MR models with $I \leq 6$ descriptors were selected (Table 7).

To obtain the best nonlinear MR model with seven descriptors, preselection of descriptors by the stepwise "*i* by *i*" procedure was used. This preselection was carried out in the same way as that for the retention times $t_R$. Thus, *i* was varied between 1 and 4, and stepwise selection was started each time from *I* descriptors ($I \leq 6$ from Table 6). The preselection procedure was halted when the models appeared containing 15−18 descriptors. In this way, 24 nonlinear MR models with 15−18 descriptors were obtained. Because some descriptors were involved in several nonlinear MR models, a novel reduced subset of 107 descriptors was obtained. From this subset, the optimum MR model with seven descriptors (Table 7) was selected, by applying the procedure for

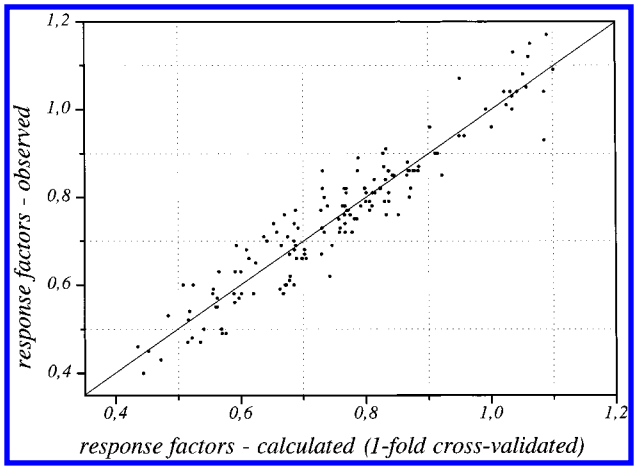


**Figure 2.** Plot of observed vs calculated response factors RF obtained using the best nonlinear seven-parameter multiregression model (model with $I = 7$ in Table 7).

selecting the best possible MR model. This model is the best of all the MR models (according to the calculated statistical parameters) for retention times obtained in this paper. In Figure 2 is given a plot of observed vs calculated (1-fold cross-validated) response factors using the seven-parameter model from Table 7.

From Figure 2 one can see that the seven-descriptor nonlinear MR model ($R_{cv}^2 = 0.9041$, $S_{cv} = 0.0501$, Table 7) is quite successful in predicting the GC response factors, although the application of the procedure described in this paper gives a smaller improvement over the CODESSA models for the prediction of GC response factors than in the case of retention times. The best seven-descriptor model from Table 7 is also tested to exclude the possibility of chance correlation. The dependent variable is randomly scrambled 1000 times, and the squares of the correlation coefficients of such RF models were between 0.006 and 0.112.

By selecting a small number of optimum descriptors, more stable regression coefficients were obtained in the MR models ($\Delta c_i$ are small comparable to the values of corresponding $c_i$ in eq 1), as can be seen from the comparison of the optimum MR models obtained in this paper with the corresponding MR models from ref 8. This result is expected, because, through selection of descriptors in "the optimum" way, the final set contains descriptors with the smallest pairwise correlation between them, which is reflected in a small standard error of estimate. Sometimes, the same optimum MR models can be accessed either through the *i* by *i* stepwise procedure or through the preselection of descriptors by their stepwise selection. This happened in the case of models with five descriptors shown in Tables 4 and 5, and in Tables 6 and 7, where the same models with five descriptors were obtained by application of "the best possible" procedure for descriptor selection, and by applying this procedure on smaller data sets which were obtained after the preselection of descriptors from the large initial data set (by the *i* by *i* stepwise selection procedure). Moreover, it is seen from Table 4 that the models with two, four, and six descriptors could be obtained by application of the 1 by 1 stepwise selection procedure applied on the models with one, three, and five descriptors from Table 4, respectively. However, it is not possible to predict whether by applying the 1 by 1 (or *i* by *i*) stepwise selection procedure on a new data set one would obtain the same models as those obtained by applying the "the best possible" way of descriptor selection.

From the comparison of the corresponding models for $t_R$ and RF in Table 2 with those in Tables 4 and 6, it is evident that all the models of Tables 4 and 6 are comparable or better. In addition, introducing the nonlinear (cross-product) terms, improved all the models, especially for modeling $t_R$. This is best illustrated if the models for $t_R$ and RF containing $I - 1$ descriptors (Tables 5 and 7) are compared with the models for $t_R$ and RF containing $I$ descriptors (Table 2). It turns out that the nonlinear MR models with $I - 1$ descriptors are better than linear ones with $I$ descriptors.

**Physical Interpretation of the Best Nonlinear Models.** The descriptors involved in the models are either similar or (often) equal to those from ref 8. Additionally, in nonlinear MR models cross-products (nonlinear terms) of initial descriptors are also involved. To interpret the meaning of such descriptors, we have to know whether the values of initial descriptors are positive or negative. All the descriptors listed in Table 3 except descriptor *d*281 (which has only negative values) have positive values for each compound.

In addition, one can see from models in Tables 4−7 that the contribution of a given descriptor in a certain model can be replaced by different combinations of several (more or less) similar descriptors in a model containing more descriptors. Conceptually this is clear, because in a multidescriptor space the same property can be represented in different ways. However, the physical meaning of some descriptors is sometimes not clear. All the initial descriptors from Table 3 involved in these models have positive values, so that their cross-products are positive. Such descriptors (obtained as products of two initial descriptors which contain all the positive values) increase/decrease the value of RT or RF if the respective regression coefficient has a positive/negative value. We obtained the best nonlinear one-descriptor model of the retention times using (cross-product) descriptor *d*137•*d*206. This is a very good descriptor and is obtained as a cross-product of *highest normal mode vibrational frequency* and *gravitation index*. Descriptor *d*137 is related to the conformational flexibility of a compound (higher conformational flexibility means higher possibility for forming stronger interaction with GC medium and longer $t_R$). As mentioned above, the gravitational index is related to the mass distribution and shape of the molecule. If both descriptors forming cross-products have higher values, $t_R$ will increase. Furthermore, we will discuss here only the physical meaning of the best nonlinear $t_R$ and RF models containing seven descriptors from Tables 5 and 7.

**(a) Physical Interpretation of the Best Nonlinear Seven-Descriptor $t_R$ Model.** Descriptors that are involved in this model are very similar to those from other previous studies on retention times.[1−8] One can identify four classes of descriptors selected for the best nonlinear seven-descriptor model: (1) descriptors describing the hydrophobicity of the molecule (*d*6, *d*9) and, thus, also hydrophobic/hydrophilic interaction between the molecule studied and the GC medium, (2) descriptors describing the ability of the molecule to form hydrogen bonds with GC medium (*d*27) and molecular solubility (*d*52), (3) descriptors representing the conformational flexibility (*d*137) and the property of the molecule in chromatographic partition process (*d*150),[38] and

(4) descriptors reflecting the size of the molecule (*d*8), the degree of branching within the molecule (*d*216), and the three-dimensional shape (*d*248, *d*252).

One of the most important descriptors is the cross-product of *the highest normal mode vibrational frequency* of the molecule (*d*137) and *the rotational entropy of the molecule at 300 K* (*d*150). Higher values of these descriptors increase retention time because the corresponding regression coefficient is positive. Compounds with a higher value of descriptor *d*137 are able to maximize favorable interactions with the GC medium. On the other hand, entering the stationary phase the molecules lose (translational, vibrational, and rotational) degrees of freedom and their entropy decreases. If the initial entropy is higher, then the interaction between the molecule and GC medium will be stronger and the energy exchanged is higher. So, a higher value of descriptor *d*137 and, at the same time, a higher value of *d*150 mean a higher retention time. The second descriptor is the number of C−X bonds (*d*8), and it models well the size of the molecule and has a positive regression coefficient in the model (the larger the molecule, the higher the retention time). Descriptor *d*9 (relative number of C−H bonds) is related to the number of CH, $CH_2$, and $CH_3$ groups and describes the ability of the molecule to participate in hydrophobic/hydrophilic interaction between the molecules and the GC medium. This descriptor is multiplied by descriptor *d*216, which describes the degree of branching of the molecule, so that the molecule with a high value of *d*9 and, at the same time, with a high *d*216 will increase retention time. Descriptor *d*6·*d*137 contains information about the molecules which can participate, at the same time, both in hydrophobic interaction and in hydrogen bonding, and a higher value of this descriptor decreases $t_R$.

**(b) Physical Interpretation of the Best Nonlinear Seven-Descriptor RF Model.** Flame ionization is a multistep process involving the thermal decomposition of a compound with subsequent "chemiionization", and the amount of ions formed determines the conductivity, which is registered as a response.[8,39] The response of hydrocarbons in the FIDs (flame ionization detectors) is attributed to the atoms from which they are made up (mainly the number of carbon atoms) and to the chemical nature of the molecules.[39] Therefore, two types of descriptors are involved in the model: (1) descriptors describing the atoms from which the molecules are constructed (*d*5, *d*9, *d*13, *d*15) and (2) descriptors which describe the chemical nature of the molecules (*d*54, *d*62, *d*63, *d*65, *d*66, *d*82; a brief description of these descriptors is given in Table 3).

The FID response of heteroatom-substituted hydrocarbons is always less than that of the parent hydrocarbon. Therefore, very important are those descriptors which contain information about the "effective" number (*d*13) or weight (*d*15) of C atoms in a compound. The "effective" carbon atom was defined as one connected only to either other carbon or hydrogen atoms.[8] In addition, the process of response of organic structures in the FID starts with the thermal decomposition of C−X bonds.[39] This can be represented by the number of different C−X bonds (where X is any atom) or by descriptors expressing the strength of such a bond, like minimal or average bond order of a carbon atom (*d*63 and *d*65) and maximal atomic orbital electronic population (*d*54). Descriptors expressing the strength of the X−H (usually C−H) bonds (minimum valency (*d*66) and atomic state energy (*d*82) of a H atom) are also very important because these bonds are on the molecular surface and they are mostly exposed in molecular collisions, so that the thermal cracking usually starts with these bonds. The descriptor which expresses the relative number of C−H bonds (*d*9) is also involved in the best RF model for the same reason.

The first three most important cross-product descriptors (*d*5·*d*82, *d*9·*d*15, *d*62·*d*66) have positive regression coefficients, and their contribution is in accordance with the physical picture given above. The descriptor *d*5·*d*82 shows that the compound with higher relative weights of C atoms and, at the same time, with higher minimum atomic state energy for a H atom produces higher response in the FID. Descriptor *d*9·*d*15 shows that the higher response is produced by those compounds with the higher relative number of C−H bonds having, at the same time, higher relative weight of effective C atoms. Descriptor *d*62·*d*66 takes into account mainly the intermolecular interactions through hydrogen bond formation. A higher value for the minimum valency of a H atom (*d*66) characterizes the compound as a weaker H-bonding donor.[8] Thus, the compound with a higher value for *d*66 having, at the same time, a higher value for the average valency of a C atom (*d*62) forms weaker hydrogen bonds between compounds, which increases the value of RF.

The dominant contribution described by the first three most important cross-product descriptors is modulated by the less important descriptors. If the square of the minimum state energy for a H atom (*d*82·*d*82) is higher, corresponding X−H bonds are stronger which makes thermal decomposition more difficult and decreases the RF value. Similarly, the higher values for both minimum and average bond orders of a C atom (*d*63·*d*65) mean stronger bonding of C atoms in the compound, which decreases the RF value. A higher atomic orbital electronic population (*d*54) and, at the same time, a higher degree of branching (*d*210) decrease the RF value.

## CONCLUDING REMARKS

New procedures for (i) the selection of descriptors (variables) in MR models, and (ii) for obtaining nonlinear MR models are successfully applied for the prediction of retention times and response factors of a very diverse set of organic compounds. The selection of descriptors is based on their orthogonalization, which enables a simple and fast computation of the correlation coefficient for a model. The set of descriptors found to achieve the highest *R* is used for the construction of the multiregression model. This algorithm allows identification of the optimum model among $10^{10}$ possible choices. For sets of possible models larger than $10^{10}$, preselection of descriptors is performed by stepwise (*i* by *i*) procedure, which provides a smaller subset of descriptors for application of the procedure for the selection of the optimum MR models. By utilizing both the linear descriptors and their cross-products, nonlinear MR models for $t_R$ and RF were obtained showing significant improvement in the statistical parameters, both fitted and cross-validated, over the linear models. Moreover, both linear and nonlinear MR models obtained by this procedure are significantly better than the models obtained by the selection of descriptors as is implemented in the CODESSA program.[8]

The best seven-descriptor nonlinear MR models for $t_R$ and RF gave the following fitted and cross-validated values of standard errors of estimate ($S = 0.3520$, $S_{cv} = 0.3743$) for $t_R$, and ($S = 0.0472$, $S_{cv} = 0.0501$) for RF. The good 1-fold cross-validated ability of these models should allow the estimation of retention times and response factors for related compounds, in the cases when experimental $t_R$ and RF values are not available.

The proposed procedure for variable selection should be applicable generally in chemometrics, analytical chemistry, medicinal chemistry, and other areas of research, where the main goal is the selection of a small subset of information-rich descriptors (variables) from a large initial data set.

In the future, it would be of great interest to compare these linear and nonlinear models with those obtained by neural networks (NN). Because MR models contain a smaller number of the optimized parameters than NN models, they should be more powerful for predicting the $t_R$ and RF for new compounds (according to Ockham's Razor[21]). Moreover, the nonlinear MR model is much simpler and enables us to inspect directly the influence of a single descriptor involved in a model on the property modeled. This was confirmed by our previous studies,[40,41] in which it was demonstrated that nonlinear MR models outperformed several NN models of biological activity of carboquinones and benzodiazepines,[25,42,43] as well as the PLS and the NN models of biological activity of antimycin analogues (Selwood data set).[22,23] Additionally, from the results obtained in the most recent comparative study[44] between MR, PLS,[45] and NN,[46−48] it also follows that the nonlinear MR produces better models.

## ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology of the Republic of Croatia through Grant 00980606. All of the computations were performed on a Hewlett-Packard 9000/E55 and SUN Enterprise 3000 at the Department of Mathematics, Faculty of Science and Mathematics, University of Zagreb. We thank Robert Manger and Vladimir Braus (Department of Mathematics, Faculty of Science, University of Zagreb, Croatia) for computer support. The authors also thank reviewers for helpful suggestions and comments.

## REFERENCES AND NOTES

(1) Georgakopoulos, C. G.; Kiburis, J. C.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Stimulants and Narcotics. *Anal. Chem.* **1991**, *63*, 2012−2024.

(2) Georgakopoulos, C. G.; Tsika, O. G.; Kiburis J. C., Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Anabolic Steroids. *Anal. Chem.* **1991**, *63*, 2025−2028.

(3) Needham, M. D.; Jurs, P. C. Quantitative Structure-Retention Relationship Studies of Polychlorinated Dibenzodioxins on Gas Chromatographic Stationary Phases of Varying Polarity. *Anal. Chim. Acta* **1992**, *258*, 183−198.

(4) Needham, M. D.; Adams, K. C.; Jurs, P. C. Quantitative Structure-Retention Relationship Studies of Polychlorinated Dibenzofurans on Gas Chromatographic Stationary Phases of Varying Polarity. *Anal. Chim. Acta* **1992**, *258*, 199−218.

(5) Woloszyn, T. F.; Jurs, P. C. Prediction of Gas Chromatographic Retention Behavior of Hydrocarbons from Naphthas. *Anal. Chem.* **1993**, *65*, 582−587.

(6) Egolf, L. M.; Jurs, P. C. Quantitative Structure−Retention Relationship and Structure−Odor Intensity Relationships for a Diverse Group of Odor-Active Compounds. *Anal. Chem.* **1993**, *65*, 3119−3126.

(7) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Alkylbenzenes. *Anal. Chim. Acta* **1997**, *342*, 113−122.

(8) 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.

(9) 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**, *24*, 279−287.

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

(11) 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.

(12) 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.

(13) 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.

(14) 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. 2. Anionic Surfactants. *J. Colloid Interface Sci.* **1997**, *187*, 113−120.

(15) 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.

(16) 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.

(17) Huibers, P. D. T.; Katritzky, A. R. Correlation of Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 283−292.

(18) Katritzky, A. R.; Sild, S.; Lobanov, V. S.; Karelson, M. Quantitative Structure−Property (QSAR) Correlation of Glass Transition Temperatures of High Molecular Weight Polymers. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 300−304.

(19) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. QSPR and QSAR Models Derived with CODESSA Multipurpose Statistical Analysis Software *J. Chem. Inf. Comput. Sci*., submitted for publication.

(20) Höskuldsson, A. Dimension of Linear Models. *Chemom. Intell. Lab. Syst.* **1996**, *32*, 37−55.

(21) Hoffmann, R.; Minkin, V. I.; Carpenter, B. K. Ockham's Razor and Chemistry. *Bull. Soc. Chim. Fr.* **1996**, *133*, 117−130.

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

(23) So, S.-S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Network. *J. Med. Chem.* **1996**, *39*, 1521−1530.

(24) Wessel, M. D.; Sutter, J. M.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants of Organic Compounds from Molecular Structure. *Anal. Chem.* **1996**, *68*, 4237−4243.

(25) Tetko, I. V.; Alessandro Villa, A. E. P.; Livingston, D. J. Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803.

(26) Lučić, B.; Nikolić, S., Trinajstić, N.; Juretić, D. The Structure−Property Models Can be Improved Using the Orthogonalized Descriptors. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 532−538.

(27) Lučić, B.; Nikolić, S.; Trinajstić, N.; Juretić, D.; Jurić, A. A Novel QSPR Approach to Physicochemical Properties of the α-Amino Acids. *Croat. Chem. Acta* **1995,** *68*, 435−450.

(28) Amić, D.; Davidović-Amić, D.; Bešlo, D.; Lučić, B.; Trinajstić, N. The Use of the Ordered Orthogonalized Multivariate Linear Regression in a Structure−Activity Study of Coumarin and Flavonoid Derivatives as Inhibitors of Aldose Reductase, *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 581−586.

(29) Lučić, B.; Trinajstić, N. New Developments in QSPR/QSAR Modeling Based on Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 45−62.

(30) Press: W. H.; Teukolsky, S. A.; Wetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge U.K., 1992.

(31) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 398 Diverse Organic and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

(32) 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.

(33) Stewart, J. J. P. *MOPAC Program Package 6.0. QCPE* **1990**, *455*.
(34) Szabo, A.; Ostlund, N. *Modern Quantum Chemistry*; McGraw-Hill: New York, 1989; pp 15−21.
(35) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure−Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.
(36) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, 97, 6609−6615.
(37) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Wiley: New York, 1986.
(38) Grunenberg, J.; Herges, R. Prediction of Chromatographic Retention Values ($R_M$) and Partition Coefficients (log $P_{oct}$) Using a Combination of Semiempirical Self-Consistent Reaction Field Calculations and Neural Networks. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 905−911.
(39) Sternberg, J. C.; Gallway, W. S.; Jones, D. T. L. The Mechanism of Response of Flame Ionization Detectors. In *Gas Chromatography*; Brenner, N., Callen, J. E., Weiss, M. D., Eds.; Academic Press: New York, 1962; pp 231−267.
(40) Lučić, B. *Quantitative Structure−Activity-Property Relationships: The Use of Non-Orthogonalized and Ordered Orthogonalized Descriptors* (in Croatian), Doctoral Thesis, Faculty of Science, University of Zagreb, Zagreb, Croatia, 1997.
(41) Lučić, B.; Trinajstić, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132.
(42) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem.* **1990**, *33*, 2583−2590.
(43) Peterson, K. L. Quantitative Structure−Activity Relationships in Carboquinones and Benzodiazepines Using Counter-Propagation Neural Networks. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 896−904.
(44) Lučić, B.; D. Amić, Trinajstić, N.; Katritzky, A. R. Multivariate Regression Outperforms Several Neural Network and Partial Least-Squares Models. Manuscript in preparation.
(45) Medven, Ž.; Güsten, H.; Sabljić, A. Comparative QSAR Study on Hydroxyl Radical Reactivity with Unsaturated Hydrocarbons-PLS versus MLR. *J. Chemom.* **1996**, *10*, 135−147.
(46) Ivanciuc, O.; Rabine, J.-P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. C-13 NMR Chemical Shift Prediction of $sp^2$ Carbon Atoms in Acyclic Alkenes Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644−653.
(47) Nefati, H.; Cense, J.-M.; Legendre, J.-J. Prediction of the Impact Sensitivity by Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 804−810.
(48) Charlton, M. H.; Docherty, R.; Hutchings, M. G. Quantitative Structure-Sublimation Enthalpy Relationship Studied by Neural Networks, Theoretical Crystal Packing Calculations and Multilinear Regression Analysis. *J. Chem. Soc., Perkin Trans. 2* **1995**, 2023−2030.

CI980161A