Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks on Three QSPR Data Sets

Bono Lučić,*,† Dragan Amić,‡ and Nenad Trinajstić†

The Rugjer Bošković Institute, P.O. Box 1016, HR-10001 Zagreb, Croatia, Faculty of Agriculture, The Josip Juraj Strossmayer University, P.O. Box 719, HR-31001 Osijek, Croatia

Received June 22, 1999

Neural networks (NNs) are accepted as the most powerful nonlinear technique in QSAR and QSPR modeling. However, the NN models are often very robust, containing a large number of parameters optimized during the training procedure. We have recently found (*J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132) that the simpler nonlinear multiregression (MR) models are significantly better than the robust NNs, according to the same statistical parameters. In the present paper we investigated whether the nonlinear MR models are also better than the concisely designed NN models. Nonlinear MR models were generated in the following way. First, nonlinear terms, the 2-fold and 3-fold cross-products of initial descriptors, were calculated and added to initial descriptors. Then, the combination of two powerful techniques for descriptor selection (*CROMRsel* for "*the best*" selection and *CROMRiisel* for approximative, "*i* by *i*" stepwise selection) were used to detect the most important descriptors in MR models. For boiling points (BPs) of 150 alkanes the 20-descriptor MR model produced the cross-validated (CV) standard error of 2.88 K, and the best NN model (with 70−80 adjusted weights) had 3.60 K. Prediction of BPs of 50 compounds using the 17-descriptor MR model (obtained on 100 compounds) gave the standard error of 3.58 K. In the case of modeling of 243 chemical shifts CV standard errors were (in ppm) 0.89 and 1.19 with 15- and 9-descriptor MR models, respectively. The best NN models adjusted 60−90 weights and achieved 1.42 ppm. The standard error in predicting the 83 chemical shifts using the 10-descriptor MR model obtained on 160 samples was 1.25 ppm. It is also shown in this data set that the model quality depends on the scaling procedure used for transformation of the initial descriptors. In modeling the sublimation enthalpy the CV correlation coefficient was 0.97 using the best 4-descriptor MR model versus 0.93 obtained using NN with ∼50 adjusted weights. The CV correlation coefficient in predicting the sublimation enthalpies for 21 compounds using the 4-descriptor MR model was 0.98. This is, to our knowledge, the first unambiguous result which shows a way for obtaining nonlinear MR models having better fitted, cross-validated, and predictive performances than the corresponding NN models. Moreover, the nonlinear MR models are significantly simpler than the NN models, which allows one to establish the functional relationships between the modeled property/activity and descriptors.

INTRODUCTION

In a related recent paper,[1] an approach for descriptor selection and multiregression (MR) model generation was described. Additionally, it has been shown that the MR models are more accurate than the models obtained by the use of several, mainly robust, neural network (NN) architectures. Robust NNs are those with $\rho < 1$ (ρ = the ratio between the number of data points and the number of connections), according to the suggestions given by the NN researchers.[2,3] It has been shown that the MR models outperform the corresponding NN models both in fitting and in cross-validation (CV) procedures. In addition, it was shown the MR models outperformed the NN models that were obtained by averaging an ensemble of 500 NNs.[1]

In the present paper we show that nonlinear multiregression can outperform even strictly defined NN architectures. Three data sets, previously studied by NN, will be used in this comparative study: (1) data set of boiling points of 150 alkanes modeled using 10 descriptors,[4] (2) data set of the ^{13}C NMR experimental chemical shifts (in ppm) of 136 alkenes corresponding 243 sp^3 carbon atoms situated in the α position relative to a double bond in acyclic alkenes (denoted as α-sp^2 carbon atoms) with distinct environments, modeled using 13 descriptors,[5] and (3) data set of sublimation enthalpies of 62 diverse organic compounds modeled using 7 descriptors.[6] According to the results obtained, Cherqaoui and Villemin[4] and Ivanciuc et al.[5] concluded that it is better to use NNs than the linear multiregression models. In the case of modeling the sublimation enthalpies, the authors[6] did not compare multivariate linear regression (MLR) and NN correctly because they did not use, in the comparison, the same statistical parameters. That is, Charlton et al.[6] used the cross-validated parameters to express the quality of the NN models and fitted parameters to measure the quality of the MLR models. In the discussion, these parameters were treated as equal.[6] Therefore, no conclusion about the quality of MR versus NN could be drawn from their work.

In all three studies the NN models were produced by multilayer back-propagation neural networks with a single

* Corresponding author. E-mail: lucic@faust.irb.hr.
† The Rugjer Bošković Institute.
‡ The Josip Juraj Strossmayer University.

10.1021/ci990061k CCC: $19.00
Published on Web 01/21/2000

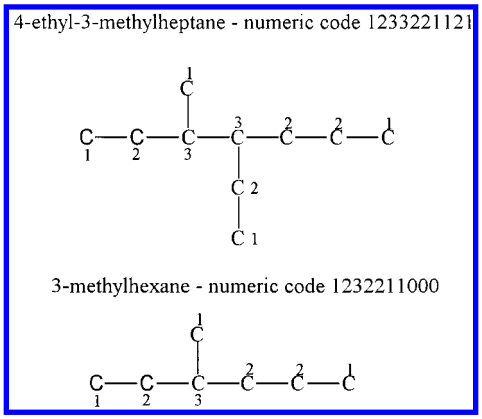

Figure 1. Examples of the numerical codes for 4-ethyl-3-methylheptane and 3-methylhexane.

hidden layer and compared with the MLR models. In modeling the boiling points only sigmoidal activation function was used, but in modeling the ^{13}C NMR chemical shifts several activation functions were used (sigmoidal, bell-shaped, symmetric logarithmoid, hyperbolic tangential, and linear) and several NN models were produced. Because all of the NNs calculated in refs 4−6 contained one hidden layer, the corresponding NN models were nonlinear,[7] and the comparison between such NN models with the linear MR models is not significant. Had the authors compared their NN models with nonlinear MR models, they would have certainly reached different conclusions. Therefore, our intention here is to show that the addition of nonlinear terms (second and third cross-product terms) to the set of initial descriptors and the use of MR-based procedure for selecting the most information-rich descriptors[1] leads to nonlinear MR models which outperform all the NN models from refs 4−6.

DATA SETS AND COMPUTATIONAL METHODS

Data Sets. The first data set (*Data set 1*) used contains the boiling points of 150 alkanes with up to 10 carbon atoms. This data set was already studied both by NN[4] and by traditional MR methods.[8] Alkanes and their boiling points (BPs) were taken from Mihalić et al.[8] Ten descriptors, which we will use, were generated in the same way as in ref 4. That is, each descriptor is given as one of the 10 digits that represent a numeric code for alkanes. Thus, one digit, which represents the number of C−C bonds in which the atom participates (e.g., the valence of a carbon atom in a hydrogen-suppressed molecular skeleton), was assigned to each carbon atom (see Figure 1).

For alkanes with number of carbons smaller than 10, the remainder of the code has the 0 values. The first digit in the numeric code, which represents the first descriptor, has the 0 value only for the methane; in all other cases it is equal to 1. Because of that, we used in this study as the nine initial descriptors digits starting from the second place in the numeric code and their nonlinear terms (2-fold and 3-fold cross-products). All descriptors used were scaled by the scaling equation (2) from ref 2 in the range between 0.1 and 1, as it was done in ref 4.

The second data set (*Data set 2*) consists of ^{13}C NMR chemical shifts of 243 α-sp^2 carbon atoms of 136 alkenes. Their structures and experimental ^{13}C NMR chemical shifts (in ppm) were taken from ref 5. According to the description

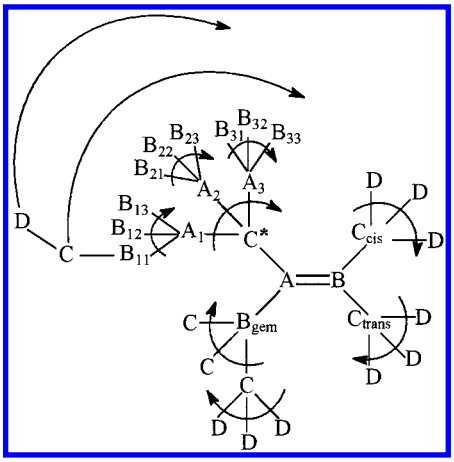

Figure 2. Description of sites in the topo-stereochemical code for the α-sp^2 carbon atoms. The resonating carbon atom is denoted by C*.

given in ref 5 (Table 2), we have generated 13 descriptors for each of 243 examples. The topo-stereochemical environment of the α-sp^2 resonating carbon atom (denoted by C* in Figure 2) is described by the neighbors denoted by A, B, C, and D positioned at 1, 2, 3, and 4 bonds away from the resonating atom. These neighbors are grouped in four spheres. The influence of the atoms positioned at a distance greater than four bonds can be neglected because they do not add much information. To take into account the stereoisomerism around the double bond, the neighboring atoms are ordered into four classes: those directly linked to C* and those linked through the gem, trans, and cis positions (see Figure 2). When the environment is ordered in the described way, it is possible to generate a topo-stereochemical (TSC) vector for each of the 243 studied cases.

The TSC vector has 13 components:

$$\mathrm{TSC}(1) = A_1 + A_2 + A_3$$

$$\mathrm{TSC}(2) = B_{11} + B_{12} + B_{13}$$

$$\mathrm{TSC}(3) = B_{21} + B_{22} + B_{23}$$

$$\mathrm{TSC}(4) = B_{31} + B_{32} + B_{33}$$

$$\mathrm{TSC}(5) = \sum C$$

$$\mathrm{TSC}(6) = \sum D$$

$$\mathrm{TSC}(7) = B_{gem}$$

$$\mathrm{TSC}(8) = \sum C_{gem}$$

$$\mathrm{TSC}(9) = \sum D_{gem}$$

$$\mathrm{TSC}(10) = C_{trans}$$

$$\mathrm{TSC}(11) = \sum D_{trans}$$

$$\mathrm{TSC}(12) = C_{cis}$$

$$\mathrm{TSC}(13) = \sum D_{cis}$$

Finally, to have the same initial descriptors as in the case of the NN modeling,[5] all the descriptors were linearly scaled between −0.9 and 0.9. In addition, the same modeling

Table 1. Details of Nonlinear Multiregression Structure−Boiling Point Models Containing *I* Descriptors Selected by SSP1 and SSP2 Procedures[a]

SSP1, all compounds			SSP2, all compounds	
descriptors	$I = 17$	$I = 20$	descriptors	$I = 18$
intercept	−256.1 (4.6)	−258.3 (3.9)	intercept	−260.5 (4.5)
1	601.5 (33.8)	598.9 (28.2)	1	640.0 (25.1)
2	38.5 (6.2)		3	264.3 (22.4)
3	192.3 (28.7)	235.0 (23.8)	4	99.0 (6.3)
4	98.6 (6.8)	85.6 (5.9)	6	46.8 (2.7)
5	66.1 (5.6)	87.3 (4.5)	1·1	−850.9 (39.1)
1·1	−668.3 (72.8)	−624.9 (62.3)	1·2	127.8 (15.2)
1·2	59.1 (7.6)	79.4 (3.8)	3·3	−342.1 (37.5)
1·3	−283.3 (78.2)	−362.0 (66.0)	5·5	120.0 (7.3)
3·6		157.5 (10.6)	1·1·1	354.9 (19.4)
6·7	81.9 (4.9)	50.0 (5.4)	1·2·2	−33.2 (10.3)
1·1·1	234.4 (43.7)	201.5 (37.8)	2·3·4	55.5 (10.5)
1·1·3	170.3 (51.3)	210.0 (43.6)	2·3·7	83.9 (4.5)
1·5·6	72.7 (6.6)		2·4·4	−77.5 (9.0)
2·2·7		23.6 (4.7)	2·8·9	18.4 (1.6)
2·4·4	−50.0 (8.3)	−43.5 (7.2)	3·3·3	156.5 (19.6)
2·7·8	54.0 (4.7)	33.5 (7.6)	5·5·5	−78.5 (7.1)
3·3·6		−172.3 (13.9)	6·7·8	46.20 (9.8)
3·3·9		55.7 (7.0)	7·8·8	23.3 (5.1)
3·5·5	−42.4 (7.2)	−57.5 (6.0)		
3·8·8		53.2 (7.6)		
3·8·9		−110.6 (18.0)		
5·6·9	−22.8 (3.8)			
8·9·9	23.7 (1.7)	42.0 (6.2)		
Statistical Parameters: This Work[b]				
R	0.9986	0.9990	R	0.9986
$R(10\%)_{cv}$	0.9980	0.9983	$R(10\%)_{cv}$	0.9980
S	2.612	2.227	S	2.555
S_{cv}	2.987	2.809	S_{cv}	3.188
$S(10\%)_{cv}$	3.124	2.881	$S(10\%)_{cv}$	3.072
statistical parameters: NN models presented in ref 4[c]				
$R = 0.9985$, $S = 2.64$, 10−7−1 configuration				
$R(10\%)_{cv} = 0.9973$, $S(10\%)_{cv} = 3.603$, 10−6−1 configuration				

[a] Numbering of descriptors is described in subsection Data Sets. Regression coefficients and their errors (in parentheses) are given for descriptors involved in MR models; I = the total number of descriptors involved in the MR model. [b] Leave-10%-out cross-validation was performed for the same training/test set partitions as it was done in ref 4. R_{cv} and S_{cv} were obtained in the leave-one-out cross-validation procedure. [c] In all the NNs 10 descriptors were used as input. Statistical parameters for the best NN model in fitting are given in Table 2 (ref 4), and for the best NN model in cross-validation described in the Prediction section of ref 4. Cross-validated parameters are calculated from the boiling points (obtained by the best NN model in cross-validation) given in Table 1 (column d) in ref 4, and are better than those calculated by authors in ref 4.

procedure was also applied to descriptors linearly scaled between −0.9 and 0.0. This was done to test the dependency of the modeling procedure on the scaling procedure, which is usually used for transformation of descriptors.

The third data set (*Data set 3*) is taken from ref 6 and contains 62 organic molecules. Each molecule was represented by seven descriptors, which were scaled between 0.1 and 1.0.

Generation of Nonlinear Multiregression Models. The multiregression models were generated mainly by the use of procedures that were described in the related papers.[1,9] Generally, the final MR models calculated in this paper are nonlinear and were obtained by applying the computer programs for the selection of the best MR models according to the highest fitted correlation coefficient (*CROMRsel*) and by computer program for the stepwise "*i by i*" selection (*CROMRiisel*). However, we performed here the stepwise selection procedure (SSP) in two different ways (SSP1 and SSP2):

(1) SSP1—stepwise selection up to K descriptors in the "*i by i*" manner ($i = 3$ in this study) starting from the best possible MR model with four descriptors selected by the use of *CROMRsel* program

(2) SSP2—stepwise selection up to K' descriptors in the "*1 by 1*" manner. In this case the stepwise selection (*CROMRiisel*) starts from each of j ($j = 1, 2, ..., N$) descriptors 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.

SSP2 is faster (takes ∼5 min, $K' = 27$) than the SSP1 (takes ∼5 h, $K = 28$) on SUN Enterprise 3000, in the case of *Data set 1* (219 descriptors). Finally, by the use of *CROMRsel* the best possible MR models with $I = 1, ..., K$ and $I = 1, ..., K'$ descriptors were selected from K and K' descriptors which were obtained by SSP1 and SSP2, respectively. For *Data sets 1* and *2* we used $K = 28$ and $K' = 27$. In the case of *Data set 3* the number of descriptors was small enough and we did not need to use the stepwise selection procedure. To obtain the models for *Data set 3*, the preselection procedure was not necessary, and all the models for this data set were obtained using only the *CROMRsel*.

In the case of *Data set 1* the nonlinearities were introduced through 2-fold cross-products of nine initial descriptors (45 descriptors), as well as 3-fold products of nine initial descriptors (165 descriptors). In *Data set 2* we had 13 initial

Table 2. Details of Three Nonlinear Multiregression Structure−^{13}C NMR Chemical Shift Models Containing *I* Descriptors Selected by the SSP1 Procedure[a]

descriptors[b]	$I = 8$	$I = 9$	$I = 15$	descriptors[c]	$I = 13$
intercept	46.86 (0.35)	47.32 (0.32)	44.45 (0.22)	intercept	38.65 (0.21)
7	5.68 (0.21)	7.02 (0.26)		1·1	−3.32 (0.28)
1·12			13.65 (0.99)	1·9	−2.17 (0.23)
1·1·2			−19.9 (1.9)	2·3	−12.08 (0.21)
1·1·8	13.7 (1.3)	14.4 (1.1)	22.1 (1.5)	1·2·3	8.36 (0.31)
1·1·12			8.8 (1.2)	1·7·8	1.38 (0.19)
1·1·13			10.08 (0.96)	1·8·12	−1.55 (0.19)
1·2·3	27.89 (0.70)	27.95 (0.63)	42.3 (1.5)	2·2·13	1.18 (0.24)
1·3·7			5.93 (0.43)	2·7·10	−0.82 (0.13)
1·5·8	−15.2 (1.2)	−15.8 (1.1)	−14.11 (0.96)	2·10·12	−1.00 (0.13)
1·8·8			−5.3 (1.1)	3·3·12	−3.46 (0.13)
1·8·12	−8.14 (0.40)	−8.22 (0.36)		3·4·8	1.67 (0.20)
1·8·13			−13.9 (1.1)	4·7·8	3.25 (0.14)
2·2·3	12.48 (0.50)	12.11 (0.45)	10.76 (0.38)	5·8·8	−3.64 (0.27)
3·10·12	−3.07 (0.37)	−3.27 (0.33)	−3.56 (0.26)		
4·7·13			5.27 (0.36)		
7·10·13		−3.60 (0.47)	−3.11 (0.35)		
8·10·13			5.17 (0.33)		
9·10·13	2.79 (0.35)	4.77 (0.40)			
R	0.9916	0.9933	0.9968	*R*	0.9938
avg $R(20\%)_{cv}$	0.9906	0.9927	0.9959	avg $R(20\%)_{cv}$	0.9921
S	1.2740	1.1366	0.7892	*S*	1.0937
avg $S(20\%)_{cv}$	1.3475	1.1918	0.8925	avg $S(20\%)_{cv}$	1.2405
min $S(20\%)_{cv}$	1.3268	1.1686	0.8668	min $S(20\%)_{cv}$	1.2073
max $S(20\%)_{cv}$	1.3885	1.2096	0.9214	max $S(20\%)_{cv}$	1.2588

[a] Numbering of descriptors is described in subsection Data Sets. Regression coefficients and their errors (in parentheses) are given for descriptors involved in MR models. I = the total number of descriptors involved in the MR model. [b] Descriptors are scaled between −0.9 and 0.0. [c] Descriptors are scaled between −0.9 and 0.9. For the models containing 11, 12, 15, and 20 descriptors, which are scaled between −0.9 and 0.9, avg $S(20\%)_{cv}$ values are 1.3630, 1.2555, 1.1141, and 0.9298, respectively.

descriptors, 91 2-fold cross-products, and 455 3-fold cross-products. Each of these nonlinear descriptors were calculated and added to initial descriptors, so that we had altogether 219 in the first and 559 descriptors in the second data set, respectively. *Data set 3* contains seven descriptors; we added to them 28 2-fold and 84 3-fold cross-products, so that the total number of descriptors in this data set was 119. Cross-products were calculated automatically from the initial data sets (this was coded as a simple subroutine).

Starting from a given initial data set of descriptors, one can decide to develop either nonlinear NN or nonlinear MR models. If one chooses to develop nonlinear MR models nonlinearities can be introduced into the MR models, in two steps: (1) calculating cross-products of initial descriptors and (2) selecting the best MR models (according to the chosen criterion) from the data set containing initial (linear) descriptors and also their cross-products using (in this case) either the SSP1 or SSP2 procedure. Final MR models have a simple additive form involving (in this study) descriptors selected among initial (linear) descriptors and their 2-fold or 3-fold cross-products. The selection procedure should be treated as a integral part of the procedure for nonlinear MR model generation.

On the other hand, NN architectures with one hidden layer automatically produce nonlinear NN models.[2] Additionally, the form and degree of nonlinearities involved into the NN models are usually unknown, but are defined by the NN architecture, by the activation function used in the hidden layer neurons, and by the learning procedure.[2]

Statistical Parameters for the Model Quality Evaluation. The quality of the fit of the models was judged in the same way as in refs 4−6, i.e., by the correlation coefficient R and standard error of estimate S:

$$S = \sqrt{\frac{\sum_{i=1}^{M}(P_i - P_i^{est})^2}{M}} \quad (1)$$

where P_i, P_i^{est}, and M denote experimental properties, estimated properties and the total number of cases (molecules) considered, respectively. These parameters were calculated for all the models. However, the more important measures of the model quality are the cross-validated (CV) statistical parameters—the correlation coefficient R_{cv} and the standard error of estimate S_{cv} (eq 2)—which are calculated using the experimental P_{cv_i} and estimated $P_{cv_i}^{est}$ properties based on the leave-one-out (S_{cv}), leave-10%-out ($S(10\%)_{cv}$), or leave-20%-out ($S(20\%)_{cv}$) cross-validation procedures.

$$S_{cv} = \sqrt{\frac{\sum_{i=1}^{M}(P_{cv_i} - P_{cv_i}^{est})^2}{M}} \quad (2)$$

In the leave-20%-out cross-validation procedure 20% of the patterns are selected at random from the complete data set to form the test (prediction) set, and remaining 80% are used for training. The model obtained in the training phase is used to predict the properties for the patterns in the test set.[10] This procedure is repeated five times, so that each of the patterns is selected only once in a test set. The leave-20%-out CV procedure was then repeated 10 times, and an average over the 10 CV values was calculated. For the best

Table 3. Details of Four Nonlinear Multiregression Structure−^{13}C NMR Chemical Shift Models Containing *I* Descriptors Selected by the SSP2 Procedure[a]

descriptors[b]	$I = 8$	$I = 9$	$I = 15$	descriptors[c]	$I = 13$
intercept	39.63 (0.42)	42.24 (0.47)	45.45 (0.31)	intercept	38.42 (0.17)
1·12		3.29 (0.66)	10.38 (1.3)	1	4.34 (0.28)
2·2	−7.71 (0.55)	−9.84 (0.58)	−11.24 (0.45)	2·3	−11.61 (0.21)
4·7	−3.21 (0.51)	−3.29 (0.50)	−4.43 (0.38)	7·8	−3.08 (0.13)
1·1·3		12.0 (1.3)	10.15 (1.1)	1·1·8	4.11 (0.29)
1·1·8	11.6 (1.3)			1·2·3	7.47 (0.33)
1·1·12			8.84 (1.2)	1·3·7	1.08 (0.19)
1·2·3	35.4 (1.3)	30.5 (1.5)	25.67 (1.2)	1·8·13	−2.13 (0.23)
1·3·7	6.07 (0.78)	5.61 (0.76)	6.35 (0.54)	2·2·13	2.27 (0.25)
1·4·5	−12.1 (1.2)	−14.3 (1.3)	−11.75 (0.93)	3·3·12	−3.71 (0.12)
1·4·9			3.93 (0.64)	3·8·10	1.18 (0.13)
1·8·8	−16.2 (1.1)	−7.57 (0.55)	−8.26 (0.60)	5·5·8	−2.69 (0.25)
1·8·12			−4.49 (0.83)	7·9·10	−0.94 (0.11)
1·11·13			2.18 (0.42)	10·12·13	−0.73 (0.12)
3·7·10			−3.78 (0.37)		
3·8·10			4.37 (0.32)		
5·10·12			−2.91 (0.30)		
12·12·12	−7.38 (0.26)	−4.69 (0.59)			
		Statistical Parameters: This Work			
R	0.9911	0.9919	0.9961	*R*	0.9941
avg $R(20\%)_{cv}$	0.9902	0.9910	0.9950	avg $R(20\%)_{cv}$	0.9929
S	1.3126	1.2502	0.8672	*S*	1.0733
avg $S(20\%)_{cv}$	1.3820	1.3186	0.9911	avg $S(20\%)_{cv}$	1.1775
min $S(20\%)_{cv}$	1.3475	1.3054	0.9622	min $S(20\%)_{cv}$	1.1507
max $S(20\%)_{cv}$	1.4143	1.3282	1.0226	max $S(20\%)_{cv}$	1.2203
		statistical parameters: NN model (the best results) presented in ref 5[d]			
R			0.9987		
avg $R(20\%)_{cv}$			0.9897		
S			0.509		
avg $S(20\%)_{cv}$			1.418		
min $S(20\%)_{cv}$			1.297		
max $S(20\%)_{cv}$			1.590		

[a] Numbering of descriptors (1−13) corresponds to TSC(1)−TSC(13) which are described in subsection Data Sets. I = the total number of descriptors involved in the MR model. Regression coefficients and their errors (in parentheses) are given for descriptors involved in MR models. [b] Descriptors are scaled between −0.9 and 0.0. [c] Descriptors are scaled between −0.9 and 0.9. For the models containing 15 and 20 descriptors, which are scaled between −0.9 and 0.9, avg $S(20\%)_{cv}$ values are 1.0257 and 0.9298, respectively. [d] Results with the best NN model in Table 4 from ref 5. In all the NNs 13 descriptors were used as input. Three factors that influence the NN performance were optimized: H = number of hidden neurons; activation function of hidden neurons Act(hidden) and output neurons Act(out). It is important to note that the fit statistical parameters of NNs (*R* and *S*) are of the NN architecture defined by Act(hidden) = bell, Act(out) = tanh, $H = 6$ (Table 3), and the cross-validated parameters are related to Act(hidden) = tanh, Act(out) = linear, $H = 6$ given in Table 4 (ref 5). These architectures produced the best results according to the given statistical parameters.

MR model of *Data set 1*, the leave-10%-out cross-validation ($S(10\%)_{cv}$, $R(10\%)_{cv}$) was performed for the same training/test set partitions as done in ref 4. In the case of *Data set 2* leave-20%-out cross-validation ($S(20\%)_{cv}$, $R(20\%)_{cv}$) with 10 randomizations was performed on the best MR models to measure more correctly the prediction quality of selected multiregression models. The same procedure was performed in ref 5. For *Data set 3* only leave-one-out CV statistical parameters were calculated in ref 6, so we also calculated only leave-one-out R_{cv} and S_{cv}.

Finally, to test the predictive capabilities of MR models, the most stringent test for evaluation of the model quality was carried out—prediction for the compounds from external data set.[10] In this case, each of the data sets was divided into training set (used for model construction) and test set (used for prediction). We used about 66% of the patterns for training and 34% for prediction. The accuracy of prediction for the molecules from the test sets was estimated by the correlation coefficient and standard error between experimental and predicted values.

Implementation. The computations were done on a Hewlett-Packard 9000/E55 (PA-RISC 7100LC processor, 100 MHz) and a SUN Enterprise 3000 (UltraSPARC processor, 250 MHz) in multiuser mode.

RESULTS AND DISCUSSION

Nonlinear MR Models on *Data Set 1*. Nonlinear MR models of boiling points of 150 alkanes are calculated starting from the total number of 219 descriptors (initial set of nine linear descriptors with 2-fold as well as 3-fold cross-products of them). Before the final nonlinear MR models were selected, the preselection of descriptors was performed using the two procedures SSP1 and SSP2. By the SSP1 procedure 28 descriptors were selected, and using the SSP2 procedure a subset of 27 descriptors was obtained. Finally, the best possible MR models (in this case the selection criterion was the highest value of R_{cv}) containing $I = 1, ..., 28$ or $I = 1, ..., 27$ descriptors were obtained using *CROMRsel* program. Details of several nonlinear MR models for boiling points of 150 alkanes that outperform corresponding NN models are given in Table 1.

For the first best NN model only fit statistical parameters are given in Table 2 in ref 4 ($R = 0.9977$, $S = 2.64$), and for the second NN model only leave-10%-out parameters

Table 4. Several the Best Possible Linear and Nonlinear Multiregression Structure−Sublimation Enthalpy Models Containing *I* Descriptors

	linear models		nonlinear models		
descriptors[a]	$I = 3$	$I = 4$	$I = 3$	$I = 4$	$I = 6$
intercept	−5.1 (1.2)	−5.14 (0.98)	4.8 (1.1)	−8.8 (1.1)	−4.5 (1.1)
1	50.1 (2.1)	61.9 (2.7)	33.4 (3.7)	60.7 (2.4)	45.1 (3.4)
5		−13.2 (2.3)			
6	20.2 (1.3)	19.6 (1.1)		19.4 (1.0)	20.15 (0.91)
7	15.1 (1.6)	16.6 (1.3)		23.1 (1.8)	
1·7					122.8 (9.7)
6·7			47.1 (2.6)		
1·1·1					52.9 (5.6)
1·1·2			26.6 (7.1)		
1·1·7					−203.7 (24.1)
1·5·7				−66.4 (10.6)	
5·5·5					−36.8 (4.5)
Statistical Parameters: This Work					
R	0.9578	0.9732	0.9588	0.9752	0.9829
R^2	0.9175	0.9470	0.9194	0.9510	0.9660
R_{cv}	0.9505	0.9648	0.9527	0.9700	0.9791
S	2.390	1.915	2.363	1.842	1.533
S_{cv}	2.585	2.187	2.529	2.024	1.696
Statistical Parameters: The Best NN Result Presented in Ref 6[b]					
R_{cv}	0.9300 ($R_{cv}^2 = 0.865$)				

[a] Descriptors 1−7: 1 = number of carbons; 2 = number of hydrogens; 3 = number of nitrogens; 4 = number of oxygens; 5 = number of π-atoms; 6 = number of hydrogen-bond donors; 7 = number of hydrogen-bond acceptors; Regression coefficients and their errors (in parentheses) are given for descriptors involved in MR models. [b] In the NN model seven descriptors were used as input.

are given in the Prediction section ($R(10\%)_{cv} = 0.9827$, $S(10\%)_{cv} = 3.70$). We recalculated these values from the data given in Table 1 in ref 4, and obtained even better values than reported: $R = 0.9985$, $S = 2.64$, $R(10\%)_{cv} = 0.9973$, $S(10\%)_{cv} = 3.60$. In Table 1 details of the models containing 17 and 20 descriptors, obtained using the SSP1 procedure, and the 18-descriptor MR model, obtained using the SSP2 procedure, are given. One can see that the models obtained using the SSP1 and SSP2 procedures outperform corresponding NN models, according to the same statistical parameters. Moreover, the presented MR models are simpler and contain a smaller number of parameters (∼20) than corresponding NN models which used 85 weights to model 150 compounds (the best model according to the fitted parameters) and 73 weights to model 135 compounds (the best CV model). One can also see that all the reported models are highly nonlinear, including many 2-fold or 3-fold cross-product terms. It should be mentioned that it is possible to obtain, for almost exactly the same data set, significantly better and simpler multiregression models involving another type of descriptors, which are better for BP modeling ($S =$ 1.92 using a linear 5-descriptor model).[11] However, our intention in this report was to compare nonlinear MR with NN models, so we had to use those data sets already studied by NN approaches, and the analysis of the descriptors quality was not our goal.

Nonlinear MR Models on *Data Set 2*. Starting from the descriptor pool containing 559 descriptors (13 initial descriptors with 2-fold and 3-fold cross-products of them), and applying the SSP1 and SSP2 procedures for preselection of descriptors, smaller subsets with 28 and 27 descriptors were preselected, respectively. Then, the best possible nonlinear MR models which contain $I = 1, ..., 28$ or $I = 1, ..., 27$ descriptors were obtained. Several selected models, for the descriptors linearly scaled both between −0.9 and 0.9 and between −0.9 and 0.0, are presented in Tables 2 and 3.

Table 5. Details of Nonlinear Multiregression Model Containing *I* Descriptors Selected by the SSP1 Procedure for the Training Set (100 Compounds) and Used for Predicting Boiling Points of 50 Compounds from the Test Set[a]

descriptors	$I = 17$, $N = 100$ (training set)[b]
intercept	−254.9 (5.5)
1	593.8 (34.2)
2	54.4 (6.1)
3	103.1 (9.3)
4	101.5 (9.8)
5	80.8 (9.5)
7	36.2 (4.2)
1·1	−714.9 (57.1)
3·5	−61.2 (12.7)
5·6	152.3 (19.7)
1·1·1	299.0 (28.7)
1·2·6	56.1 (13.7)
1·3·6	−56.0 (18.5)
2·3·9	20.5 (2.2)
2·4·4	−62.3 (12.2)
2·7·8	52.1 (9.4)
5·5·6	−109.9 (18.5)
8·8·8	10.6 (2.3)
Statistical Parameters: In Training[b]	
R[c]	0.9985
S	2.862
Statistical Parameters: In Prediction[b]	
R(test set)[c]	0.9966
S(test set)	3.598

[a] See footnote *a* in Table 1. [b] In the prediction procedure, the data set is split into the training set containing 100 compounds and the test set containing 50 compounds. The best 17-descriptor MR model, obtained on the training set, is used for prediction on the test set. The test set compounds are as follows (the numbering corresponds to that in ref 4): 5, 7, 9, 11, 13, 15, 17, 19, 23, 29, 31, 33, 39, 43, 47, 49, 51, 55, 57, 59, 61, 63, 65, 67, 77, 79, 81, 87, 89, 91, 93, 95, 103, 105, 107, 109, 111, 115, 117, 119, 121, 125, 127, 131, 135, 137, 139, 141, 145, 149, and the remaining 100 compounds are involved in the training set. [c] R and S measure the fit performance of the model obtained on the training set, and R(test set) and S(test set) (root-mean-square error between experimental and predicted BPs) measure the quality of the model in prediction for 50 compounds from the test set.

Table 6. Details of Three Nonlinear Multiregression Models Containing *I* Descriptors Selected by the SSP1 Procedure for the Training Sets (160 Carbon Atoms) and Used for Predicting ^{13}C NMR Chemical Shifts of 83 Carbon Atoms from the Test Sets[a]

partition-1: $N=160$ (training set)[b]				partition-2: $N=160$ (training set)[b]			
$I=13$ ($-0.9-0.0$)[c]		$I=16$ ($-0.9-0.9$)[d]		$I=13$ ($-0.9-0.0$)[c]		$I=16$ ($-0.9-0.9$)[d]	
intercept	42.31 (0.62)	intercept	36.64 (0.24)	intercept	46.50 (0.37)	intercept	37.10 (0.23)
7·7	−5.98 (0.40)	12	−2.98 (0.14)	7·7	−7.04 (0.44)	2·3	−13.66 (0.23)
10·12	3.83 (0.33)	1·9	−3.12 (0.24)	1·1·7	3.17 (0.62)	5·6	4.35 (0.31)
1·1·7	4.24 (0.57)	2·3	−14.76 (0.4)	1·1·8	6.2 (1.2)	1·1·8	3.40 (0.25)
1·1·9	12.46 (0.97)	4·5	4.59 (0.38)	1·1·12	12.4 (1.3)	1·1·12	1.45 (0.25)
1·2·3	28.40 (1.05)	1·1·8	1.62 (0.23)	1·2·3	26.27 (0.81)	1·2·6	8.61 (0.39)
1·5·5	−9.08 (0.77)	1·1·11	2.68 (0.27)	1·5·12	−13.72 (1.4)	1·2·8	−1.45 (0.25)
1·8·9	−9.16 (0.79)	1·1·12	1.12 (0.24)	1·8·8	−9.1 (1.0)	1·4·12	−1.70 (0.19)
1·8·12	−7.70 (0.36)	1·2·3	7.34 (0.29)	1·8·12	−5.32 (0.83)	1·6·6	3.50 (0.39)
2·2·3	11.47 (0.50)	1·7·8	1.62 (0.17)	1·10·12	−5.43 (0.62)	1·7·8	1.46 (0.17)
3·7·10	−3.13 (0.40)	1·8·12	−1.29 (0.19)	1·10·13	3.63 (0.67)	2·2·13	2.90 (0.26)
4·9·10	2.36 (0.47)	2·2·3	3.20 (0.34)	2·2·3	11.76 (0.45)	2·8·13	−2.51 (0.25)
8·10·13	2.93 (0.47)	2·8·11	−1.54 (0.21)	4·9·10	2.72 (0.45)	3·3·12	−3.85 (0.18)
9·9·9	−4.43 (0.82)	2·9·9	−3.20 (0.43)	5·7·10	−3.94 (0.46)	7·8·9	3.12 (0.14)
		3·7·10	−0.55 (0.10)			7·8·10	−1.01 (0.12)
		7·8·9	2.82 (0.13)			10·11·11	0.81 (0.14)
		10·11·12	−0.93 (0.12)			10·12·13	−0.65 (0.13)
Statistical Parameters: In Training[b]							
R[c]	0.9967		0.9970		0.9959		0.9963
S	0.831		0.796		0.893		0.851
Statistical Parameters: In Prediction[b]							
R(test set)[e]	0.9910		0.9893		0.9936		0.9939
S(test set)	1.269		1.412		1.119		1.083

[a] See footnote *a* in Table 1. [b] In the prediction procedure, the data set is split into the training set containing 160 carbon atoms and the test set containing 83 carbon atoms. The best 13-descriptor MR model, obtained on the training set, is used for prediction on the test set. The test sets carbon atoms are as follows (the numbering corresponds to that in ref 5): 2, 6, 8, 10, 12, 16, 20, 22, 24, 36, 38, 40, 42, 44, 48, 50, 52, 56, 60, 62, 64, 68, 70, 74, 78, 80, 82, 84, 86, 90, 92, 94, 98, 100, 102, 104, 112, 114, 118, 122, 124, 128, 130, 132, 134, 136, 142, 146, 148, 152, 154, 158, 162, 164, 166, 168, 172, 174, 176, 178, 182, 184, 186, 188, 192, 196, 198, 202, 204, 208, 210, 212, 214, 216, 218, 222, 224, 226, 230, 236, 238, 240, 242 (partition-1); and 4, 5, 6, 8, 10, 18, 21, 22, 25, 27, 28, 31, 32, 34, 45, 47, 50, 57, 60, 64, 66, 69, 70, 71, 72, 74, 81, 86, 88, 91, 93, 98, 99, 101, 107, 112, 118, 120, 124, 126, 128, 132, 133, 134, 144, 152, 153, 157, 161, 165, 166, 167, 168, 170, 172, 176, 181, 183, 185, 186, 190, 195, 197, 200, 201, 203, 205, 210, 212, 214, 216, 219, 220, 223, 224, 225, 226, 231, 232, 233, 236, 239, 243 (partition-2). [c] Descriptors are scaled between −0.9 and 0.0. [d] Descriptors are scaled between −0.9 and 0.9. [e] R and S measure the fit performance of the model obtained on the training set, and R(test set) and S(test set) (root-mean-square error between experimental and predicted values) measure the quality of the model in prediction for 83 carbon atoms from the test set.

Nonlinear MR models involving descriptors scaled between −0.9 and 0.9 that are better than the best NN models (according to fitted and cross-validated parameters) contain 13 or more descriptors (see Tables 2 and 3 and footnotes *c* in them). In addition, one can see that even the nonlinear MR models with only eight descriptors (scaled between −0.9 and 0.0) obtained by the SSP1 (Table 2) and SSP2 (Table 3) procedures have better performances than the best NN models given in ref 5. Moreover, the nonlinear 7-descriptor MR model obtained by the SSP1 (eq 3) has a performance almost equal to that of the best NN models:

$$\begin{aligned}{}^{13}\text{C NMR shift} = {} & (44.562 \pm 0.333) + (3.918 \pm 0.443)\text{d7} + (18.930 \pm 1.211)\text{d1·d12} + (13.719 \pm 1.519)\text{d1·d1·d12} + (32.987 \pm 0.771)\text{d1·d2·d3} + \\ & (3.896 \pm 0.756)\text{d1·d3·d7} + (-8.250 \pm 0.628)\text{d1·d5·d8} + (9.850 \pm 0.510)\text{d2·d2·d3} \quad (3)\end{aligned}$$

$$N = 243, \quad R = 0.9906, \quad \text{avg}R(20\%)_{\text{cv}} = 0.9896, \quad S = 1.3518, \quad \text{avg}S(20\%)_{\text{cv}} = 1.4224$$

In this equation descriptors TSC(1), ..., TSC(13) are denoted as d1, ..., d13, and in Tables 2 and 3 simply as 1, ..., 13. One can see that the most important nonlinear descriptors in eq 3 involve linear descriptors d1 (type A neighbor), d2 and d3 (type B neighbors), which describe the adjacent environment of the resonant C* atom. Linear descriptors do not appear to play any significant role in competition with the nonlinear combinations. Insertion of a carbon atom at a certain position affects the magnetic field at C* not only directly, but also indirectly, via interactions with other neighboring carbon atoms. Such a phenomenon is better taken into account by an appropriate nonlinear descriptor.

From the physical considerations it is difficult to predict the sign and magnitude of the chemical shift induced by the presence of various A, B, C, and D carbon atoms. The dominating contribution to the C* shift is the dielectric term originating from the excitation of the electrons of the sp^3 hybrid orbitals of the carbons adjacent to C* (descriptor d1). Thus, the nonlinear descriptors expressing the dominating role of d1 are also expected to give rise to the positive chemical shift (descriptors d1·d2·d3, d1·d3·d7, d1·d12, d1·d1·d3, etc). The prediction of the sign of more remote carbons, particularly in the nonlinear descriptors, is practically impossible.

Low frequency or even the lack of some linear descriptors (d4, d5, d6, d9, d11, and d13) in the nonlinear combinations is only partly due to the fact that they contain the information of the remote carbon atoms. In addition, descriptors d6 and d11 are not included in the models given in Tables 2 and 3. It is also true that the set of data (*Data set 2*) used for the model and the testing is highly asymmetric in the appearance of the carbon atoms of class B. From 243 data points, B_1 is

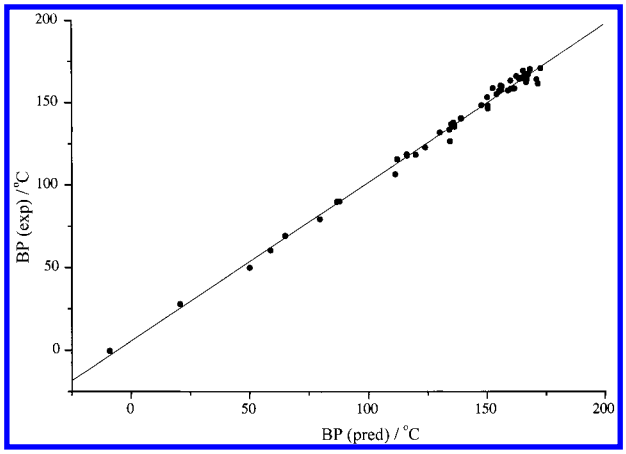

Figure 3. Plot of the experimental versus predicted BP values (in °C) of 50 alkanes from the external set.

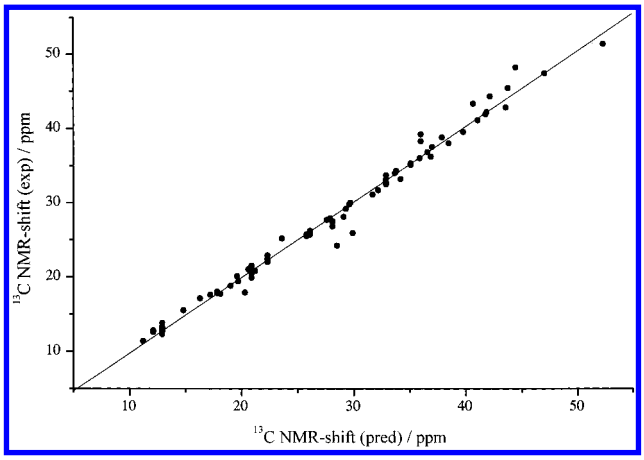

Figure 4. Plot of the experimental versus predicted ^{13}C NMR chemical shifts (in ppm) of alkenes from the external set containing 83 values (partition-1), using the 13-descriptor model from Table 6.

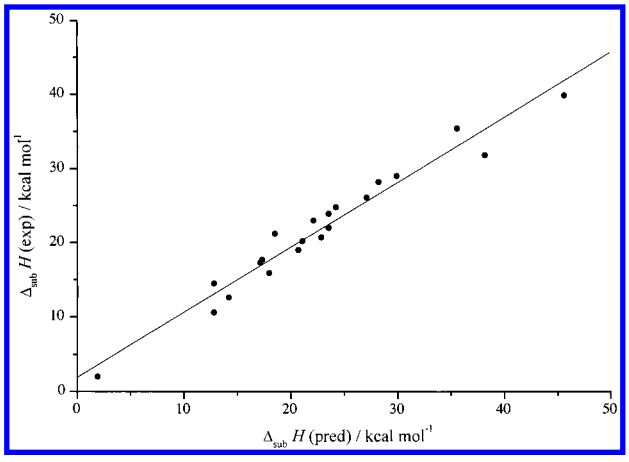

Figure 5. Plot of the experimental versus predicted sublimation enthalpy ($\Delta_{sub}H$/kcal mol^{-1}) values of 21 organic compounds from the external set.

populated in 155 cases, B_2 in 56 cases, and B_3 only in 27 cases. Thus, it is quite reasonable to expect a relatively low statistical weight of the B_3 carbons and hence of descriptor d4.

Considering the MR models with more descriptors, we see that such models are much better than NN models, according to the same statistical parameters. The most surprising fact is that the differences between fitted and CV

Table 7. Details of the Best Linear Multiregression Models for the Training Set (41 Compounds) Used for Predicting Sublimation Enthalpies of 21 Compounds from the Test Set

descriptors[a]	$I = 3$, $N = 41$ (training set)[b]	$I = 4$, $N = 41$ (training set)[b]
intercept	3.3 (1.2)	2.84 (0.93)
1	1.456 (0.077)	1.89 (0.10)
5		−0.461 (0.089)
6	4.73 (0.38)	4.57 (0.29)
7	2.17 (0.31)	2.58 (0.25)
Statistical Parameters: In Training[b]		
R	0.9559	0.9749
R_{cv}[c]	0.9452	0.9678
S	2.430	1.843
Statistical Parameters: In Prediction[b]		
R(test set)	0.9639	0.9781
R(test set)^2	0.9290	0.9567
S(test set)	2.391	2.311

[a] See footnote *a* in Table 4. [b] The training set contains 41 compounds and the test set 21 compounds. The best *I*-descriptor MR model, obtained on the training set, is used for prediction on the test set. The test set compounds are as follows (the numbering corresponds to that in ref 6): 2, 3, 7, 9, 15, 16, 26, 28, 30, 33, 38, 40, 41, 43, 45, 47, 48, 55, 56, 60, 62. [c] In this case the leave-one-out cross-validation procedure was performed.

statistical parameters are much smaller for MR than for NN models. In addition, MR models presented in Tables 2 and 3 contain a smaller number of adjusted parameters (9−16) than the best NN models (61−91 weights). Application of different linear scaling procedures leads to the models of different quality containing different type and number of descriptors. This is caused by nonlinear terms (cross-products) used in modeling. That is, differently linearly scaled initial (linear) descriptors give different nonlinearly scaled cross-products (i.e., linear scaling of initial descriptors leads to nonlinear scaling of cross-product descriptors).[12]

Linear and Nonlinear MR Models on *Data Set 3*. Modeling on this data set started from seven descriptors for 62 compounds that were taken from ref 6 and scaled between 0.1 and 0.9. To these initial (linear) descriptors their 2-fold and 3-fold cross-products were added; thus, this data set contains 119 descriptors. On this data set we applied only the procedure for selecting the best possible MR models with up to six descriptors. Statistical parameters for the best nonlinear MR models and the best 3- and 4-descriptor linear MR models are given in Table 4.

It is evident (by means of the calculated statistical parameters) that several MR models obtained for this data set are much better than the NN models calculated in ref 6. Moreover, the difference between NN and MR models for this data set are most pronounced among all three data sets studied here. We can see from Table 4 that by including nonlinear terms the MR models are only slightly improved. From this fact we can conclude that the relationships between sublimation enthalpy and molecular structure (which is coded by the calculated descriptors) is mostly linear. In addition, this also shows that some NN architectures (together with some procedures used for their training) are not able to model even linear relationships.

Predictive Performances of the MR Models. In addition to the comparison described above, we also tested the predictive capabilities of nonlinear MR models. Because such a study was not originally undertaken in the reported NN

Table 8. Predicted Values of Boiling Points, ^{13}C NMR Chemical Shifts, and Sublimation Enthalpies for the Test Set Compounds or Carbon Atoms Using the Multiregression Models Given in Tables 5−7[a]

boiling points prediction by 13-desciptor-model from Table 5′ compds[a]/exp/prd	^{13}C NMR chemical shifts		sublimation enthalpies prediction by 4-descriptor model from Table 7 compds[d]/exp/prd
	partition-1: prediction by 13- and 16-descriptor models from Table 6 C atoms[b]/exp/prd1/prd2[c]	partition-2: prediction by 13- and 16-descriptor models from Table 6 C atoms[b]/exp/prd1/prd2[c]	
5/−0.50/−9.13	2/24.2/25.5/24.8	4/11.4/11.2/12.4	2/12.6/14.2
7/27.80/20.62	6/25.5/26.3/24.3	5/17.1/16.3/17.6	3/15.9/18.0
9/49.70/50.02	8/20.4/20.7/20.0	6/25.5/25.8/26.2	7/17.3/17.1
11/60.30/58.79	10/31.1/31.2/32.2	8/20.4/20.9/21.2	9/20.7/22.9
13/69.00/64.96	12/25.8/25.7/25.9	10/31.1/31.7/31.3	15/26.1/27.1
15/79.20/79.70	16/32.6/31.9/31.6	18/13.3/12.9/12.5	16/23.9/23.6
17/89.80/86.84	20/22.9/22.2/21.5	21/13.0/12.9/12.9	26/19.0/20.7
19/90.00/87.88	22/21.5/22.0/21.5	22/21.5/20.9/21.0	28/2.0/1.9
23/106.50/111.39	24/25.7/26.3/25.3	25/27.7/27.6/27.7	30/10.6/12.8
29/118.20/120.04	36/20.7/21.1/20.9	27/20.6/20.9/21.0	33/14.5/12.8
31/115.60/112.24	38/13.1/13.8/14.9	28/19.9/20.9/20.6	38/17.7/17.3
33/117.70/116.45	40/28.5/29.8/30.1	31/20.1/19.6/20.5	40/20.2/21.1
39/118.50/116.28	42/12.4/12.8/12.7	32/28.1/29.1/28.8	41/24.8/24.2
43/131.70/130.31	44/15.7/15.8/16.4	34/12.3/12.9/13.4	43/29.0/29.9
47/137.70/136.02	48/21.2/22.0/21.5	45/21.3/20.9/20.5	45/35.4/35.6
49/122.70/124.12	50/17.7/18.3/17.9	47/22.9/22.3/22.7	47/23.0/22.1
51/126.50/134.57	52/25.2/26.7/25.1	50/17.7/18.1/18.6	48/28.2/28.2
55/135.20/136.58	56/19.6/15.4/16.6	57/34.3/33.8/33.3	55/31.8/38.2
57/146.20/150.64	60/18.0/18.4/20.0	60/18.0/17.8/18.1	56/22.0/23.6
59/136.73/135.08	62/14.3/13.3/12.2	64/25.7/26.1/25.7	60/21.2/18.5
61/140.50/139.19	64/25.7/25.7/25.6	66/20.8/21.2/21.4	62/39.8/45.7
63/140.10/139.27	68/22.4/22.8/20.9	69/13.8/12.9/12.4	
65/133.50/134.34	70/29.2/29.8/28.1	70/29.2/29.3/27.9	
67/136.00/136.58	74/18.8/19.0/17.9	71/24.2/28.5/26.0	
77/158.00/156.36	78/12.4/12.8/13.1	72/12.7/12.9/12.3	
79/170.50/173.04	80/31.6/30.4/31.0	74/18.8/19.0/19.1	
81/154.90/154.40	82/32.6/32.3/34.2	81/27.9/27.9/28.1	
87/164.00/167.31	84/33.2/31.5/32.2	86/38.3/36.0/37.1	
89/164.59/164.57	86/38.3/37.8/36.5	88/39.2/36.0/37.9	
91/170.00/168.61	90/19.5/14.0/15.5	91/36.2/36.9/36.7	
93/153.00/150.45	92/40.3/40.4/41.1	93/22.4/22.3/22.0	
95/169.00/165.64	94/35.0/35.0/34.9	98/42.2/41.9/42.2	
103/158.00/160.32	98/42.2/41.2/40.6	99/15.5/14.8/15.1	
105/148.00/150.56	100/13.3/12.8/12.8	101/33.7/32.9/33.3	
107/148.20/147.95	102/23.4/21.5/20.7	107/36.8/36.6/36.9	
109/161.20/171.91	104/30.4/31.2/30.4	112/44.3/42.2/41.8	
111/163.80/171.37	112/44.3/41.8/42.6	118/33.2/34.2/33.9	
115/163.00/160.28	114/38.7/38.4/39.1	120/35.1/35.1/34.9	
117/164.00/164.02	118/33.2/33.9/34.1	124/45.4/43.8/43.5	
119/157.04/159.19	122/29.6/30.4/29.9	126/13.2/12.9/12.5	
121/157.00/155.51	124/45.4/42.5/42.0	128/17.9/20.3/20.7	
125/164.31/165.83	128/17.9/18.1/18.8	132/19.4/19.7/20.0	
127/166.00/166.20	130/36.7/38.5/39.6	133/43.3/40.7/40.9	
131/167.00/167.89	132/19.4/16.7/17.6	134/27.5/28.1/26.7	
135/160.00/156.22	134/27.5/26.3/26.1	144/37.5/37.0/37.5	
137/158.30/161.98	136/39.1/36.9/34.2	152/29.8/29.6/29.6	
139/159.70/156.66	142/36.3/37.0/36.7	153/42.8/43.6/42.7	
141/158.54/152.76	146/17.8/17.5/17.9	157/25.2/23.6/21.0	
145/165.70/162.77	148/47.9/44.9/46.5	161/31.7/32.2/32.6	
149/162.00/166.90	152/29.8/30.4/29.9	165/48.2/44.5/49.0	
	154/20.8/19.5/20.6	166/38.8/37.9/42.5	
	158/44.7/44.8/42.6	167/51.4/52.4/52.1	
	162/20.0/19.9/20.1	168/47.4/47.1/48.5	
	164/34.3/32.4/33.8	170/32.6/32.9/33.1	
	166/38.8/37.1/35.6	172/26.8/28.1/27.2	
	168/47.4/46.7/47.4	176/32.8/32.9/32.6	
	172/26.8/27.2/27.4	181/17.6/17.2/18.1	
	174/37.8/38.0/38.6	183/41.1/41.1/39.4	
	176/32.8/32.5/32.7	185/36.0/35.9/36.6	
	178/27.3/27.9/27.4	186/30.0/29.7/28.2	
	182/25.7/26.3/25.3	190/35.3/35.1/34.9	
	184/18.9/19.4/19.4	195/32.5/32.9/32.6	
	186/30.0/27.8/27.7	197/34.0/33.7/34.7	
	188/41.3/41.4/41.4	200/27.4/28.1/27.6	
	192/29.9/30.4/29.9	201/12.6/12.1/12.9	
	196/27.5/27.9/27.4	203/25.8/26.1/25.7	
	198/33.2/32.5/32.4	205/21.0/20.6/20.9	
	202/32.8/32.5/32.7	210/25.9/29.9/28.2	
	204/27.7/27.9/27.4	212/25.7/25.8/25.7	

Table 8 (Continued)

boiling points prediction by 13-desciptor-model from Table 5′ compds[a]/exp/prd	^{13}C NMR chemical shifts		sublimation enthalpies prediction by 4-descriptor model from Table 7 compds[d]/exp/prd
	partition-1: prediction by 13- and 16-descriptor models from Table 6 C atoms[b]/exp/prd1/prd2[c]	partition-2: prediction by 13- and 16-descriptor models from Table 6 C atoms[b]/exp/prd1/prd2[c]	
	208/27.8/27.9/27.4	214/22.4/22.3/22.0	
	210/25.9/27.3/25.4	216/17.8/17.8/17.5	
	212/25.7/26.3/25.3	219/33.2/32.9/33.0	
	214/22.4/22.0/21.2	220/26.2/26.1/25.7	
	216/17.8/18.4/18.0	223/32.8/32.9/33.5	
	218/12.7/11.9/12.2	224/17.9/17.8/17.5	
	222/21.1/21.1/20.9	225/27.5/28.1/27.6	
	224/17.9/18.4/18.0	226/12.8/12.1/12.9	
	226/12.8/11.9/12.2	231/38.0/38.5/38.5	
	230/34.0/34.0/34.5	232/22.0/22.3/22.0	
	236/34.0/34.0/34.5	233/34.0/33.7/34.7	
	238/42.1/41.7/38.7	236/34.0/33.7/34.7	
	240/42.0/41.7/41.2	239/39.5/39.8/38.2	
	242/41.9/41.7/43.7	243/41.9/41.8/42.3	

[a] The numbering corresponds to that in ref 4. [b] The numbering corresponds to that in ref 5. [c] prd1: prediction is done using the 13-descriptor model (descriptors are scaled between −0.9 and 0.0); prd2: prediction is done using the 16-descriptor model (descriptors are scaled between −0.9 and 0.9). [d] Position of the compounds according to Table 1 from ref 6.

works,[4−6] we cannot use the obtained results to compare predictive performances of MR with those of NN models. However, these results should be indicators for the predictive capabilities of the MR models alone (testing of the models in prediction, i.e., external validation, is the strongest procedure for expressing the real model quality[10]). In all cases the models are generated following the same procedures as described above. Details of the models developed on the training sets which were used in prediction, and their statistical parameters, are given in Tables 5, 6, and 7 for *Data set 1*, *2*, and *3*, respectively.

Prediction test for *Data set 1* was performed on 50 compounds using the model obtained on the remaining 100 compounds. The nonlinear 17-descriptor MR model was generated on the training set using the SSP1 procedure (see Table 5).

This MR model achieved a lower standard error in prediction than the NN model in cross-validation. From Figure 3 one can see that the prediction of boiling points for 50 compounds is very good in all ranges of experimental values.

In this case, as well as in two next cases (Figures 4 and 5) the straight line represents the linear regression of the data, from which the correlation coefficients and standard errors of prediction are calculated.

In the case of *Data set 2*, the model was constructed on 160 carbon atoms and then ^{13}C NMR chemical shifts for the remaining 83 carbon atoms was predicted. To better evaluate the modes quality, two training/set partitions were performed, and in both cases very good predictions were obtained. Molecules involved in the test sets are given in Table 6, as well as the statistical parameters, descriptors, and corresponding regression coefficients.

The overall standard errors of prediction obtained using nonlinear 13-descriptor MR models (using descriptors scaled between −0.9 and 0.0) for the test set carbon atoms are 1.269 and 1.119 ppm for partition-1 and partition-2, respectively. Standard errors of prediction obtained by nonlinear 16-descriptor MR models from Table 6 (using descriptors scaled between −0.9 and 0.9) are 1.412 and 1.083 ppm for partition-1 and partition-2, respectively. These values of standard errors are even lower than the cross-validated values obtained by the best NN models from ref 5.

Data set 3 was split into a training set which contained 41 compounds and a prediction set with 21 compounds. Details of the descriptors involved in the MR models obtained on the training set, the corresponding multiregression coefficients, and statistical parameters for the 3- and 4-descriptor models are given in Table 7.

The standard error of prediction of sublimation enthalpy is very close to corresponding fitted and cross-validated parameters given before. The correlation between predicted and experimental values for *Data set 3* is shown in Figure 5.

The lists of compounds or carbon atoms used for prediction (we give their position according to tables in refs 4−6) and the corresponding experimental and predicted values for boiling points, ^{13}C NMR chemical shifts, and sublimation enthalpies are given in Table 8.

It is important to note that only the best nonlinear MR models obtained by either the SSP1 or SSP2 procedures are given in Tables 1−7. Applying the SSP1 and SSP2 selection procedures, we selected several MR models having almost the same fitted, cross-validated, and predictive performances as the top MR models from Tables 1−7, but still better performances than the best NN models given in refs 4−6. This further favors MR-based selection procedures and models over NN approaches.

CONCLUSION

Using the statistical parameters, which are exactly those originally used in the NN analyses,[4−6] we showed that, in three studied cases, the nonlinear MR models outperformed the corresponding NN models. Moreover, we have shown that we can select, for each data set, not only one but also several nonlinear MR models that are better than the best NN models obtained. Such results are in agreement with Ockham's razor,[13] which prefers the models realized with the fewest descriptors, other things being equal.

Additionally, such models are very good in prediction on the test set compounds or carbon atoms (which were not used for model generation). Their standard errors in prediction were better than CV standard errors of the corresponding NN models. It was also shown that the quality of the models is dependent on the scaling procedure used for transformation of descriptors. In comparative studies this fact should be taken into account, to have an exact comparison between different methods.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology of the Republic of Croatia through Grants 00980606 (N.T. and B.L.), 106N407 (B.L.), and 079301 (D.A.). We gratefully acknowledge Professor Janko Herak (Faculty of Pharmacy and Biochemistry, University of Zagreb, Croatia) for his help and discussions about the physical meaning of descriptors involved in the models of the ^{13}C NMR chemical shifts. We thank Robert Manger and Vladimir Braus (Department of Mathematics, Faculty of Science, University of Zagreb, Croatia), and Ljubimko Šimičić (Department of Physics, Faculty of Science, University of Zagreb, Croatia) for computer support. The authors thank the reviewers for their helpful remarks.

REFERENCES AND NOTES

(1) 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.

(2) 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.

(3) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.

(4) Cherqaoui, D.; Villemin, D. Use of a Neural Networks to Determine the Boiling Points of Alkanes. *J. Chem. Soc., Faraday Trans.* **1994**, *90*, 97−102.

(5) Ivanciuc, O.; Rabine, J.-P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. ^{13}C NMR Chemical Shift Prediction of the sp^3 Carbon Atoms in the α Position Relative to the Double Bond in Acyclic Alkenes. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 587−598.

(6) Charlton, M. H.; Docherty, R., Y.; 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.

(7) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem*. **1990**, *33*, 2583−2590.

(8) Mihalić, Z.; Nikolić, S.; Trinajstić, N. Comparative Study of Molecular Descriptors Derived from the Distance Matrix. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 28−37.

(9) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 610−621.

(10) Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of Comparative Molecular Field Analysis Models: Effect of Data Scaling and Variable Selection Using a Set of Human Synovial Fluid Phospholipase A2 Inhibitors. *J. Med. Chem*. **1997**, *40*, 1136−1148.

(11) Rücker, G.; Rücker, C.; On Topological Indices, Boiling Points and Cycloalkanes. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 788−802.

(12) After linear scaling of descriptors d1 and d2 we obtained descriptors $(d1/a_1) - b_1$ and $(d2/a_2) - b_2$, where a_1, b_1, a_2, and b_2 are constants. Multiplying scaled descriptors give us a new cross-product descriptor $d3 = d1 \cdot d2/(a_1 \cdot a_2) - d1 \cdot b_2/a_1 - d2 \cdot b_1/a_2 + b_1 \cdot b_2$. Thus, for differently linearly scaled descriptors d1 and d2 (with different a_1, b_1, and a_2, b_2), we obtain differently, and nonlinearly scaled, cross-product descriptor d3.

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

CI990061K