Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling

Bono Lučić*,† and Nenad Trinajstić‡

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

Received May 27, 1998

In the past decade, many authors replaced multivariate regression (MR) by the neural networks (NNs) algorithm because they believed the latter to be superior. To verify this, we have undertaken a comparative investigation of the relationship between biological activities and substituent constants representing physicochemical parameters of the substituent groups of 37 carboquinones and 57 benzodiazepines using MR and NNs. A new method for the selection of descriptors in the best possible MR models is presented. The use of orthogonalization procedure makes the calculation of the statistical parameters (e.g. correlation coefficient, *R*) for each model much simpler, and the selection of the best MR models is accelerated. Such a procedure is applicable to QSAR modeling even for the selection of the best MR model with six descriptors from a set of 100 descriptors. In case one wants to select, for example, the best 15 out of 100 descriptors, a new procedure is developed for the stepwise selection of descriptors in MR models. Using this procedure, we selected not only one (which was the case in the old stepwise MR procedure) but two, three, or more new descriptors in each subsequent step and added them to descriptors selected up to the previous step. The same data sets were previously investigated by several (mainly robust) NN algorithms which contained a hidden layer (Aoyama, T. et al. *J. Med. Chem.* **1990**, *33*, 2583−2590; Peterson, K. L. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 896−904; Tetko, I. V. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803), and the authors have concluded that NNs models are better than MR models. These NNs are mainly robust, i.e., contain a large number of connections, and consequently, there are many parameters (weights) that should be optimized. Since it is well-known that NNs with hidden layers take into account nonlinear operations, for a strict comparison between NNs and MR the initial descriptors set used for obtaining MR models should include also nonlinearities. This was done by enlarging the initial descriptor set by including squares and cross-products of initial descriptors. After that, a systematic comparison between MR and this specific architectures of NNs was carried out on seven QSAR models, and MR models were superior in all studied cases.

INTRODUCTION

The fundamental axiom of QSAR (quantitative structure−activity relationship) is that biological properties of molecules depend on their structures.[1,2] QSAR is an important tool in molecular design, especially in developing new drugs.[3,4] The first step in QSAR modeling is the representation of molecular structures by numbers. These numbers are called molecular descriptors.[5] They are commonly computed by using quantum-chemical, information-theoretic, graph-theoretic, or geometric approaches.[6−9] Computer programs that can compute a variety of descriptors (up to 400 descriptors) are available.[10,11]

After computing molecular descriptors, a procedure has to be established that can best utilize descriptors in QSAR modeling. Depending on the mathematical approach used in the analysis, the final QSAR models may be quite different in their complexity, accuracy, stability, and predictability. The QSAR approach has evolved from a simple regression model with a few variables[12] to an important tool applicable to a wide range of chemical, biological, and medicinal problems.[13−15] In the past 30 years many new techniques have been introduced for obtaining QSARs, for which it is actually believed that they go significantly beyond the original linear (nonlinear) regression analysis.[16] The most widely used techniques are the principal component analysis (PCA),[17] partial least squares (PLS),[18,19] genetic algorithms (GA),[20] and neural networks (NNs).[21,22] Moreover, a hybrid approach has been recently published as a combination of genetic algorithm and neural networks, and it was shown that this GA-NN method is the most successful method in QSAR modeling (according to statistical parameters that authors calculated).[16] Besides in medicinal chemistry (structure−activity and structure−property relationships), all these methods are also widely used in many areas of chemometrics and analytical chemistry (spectroscopy, development of chemical instrumentation, optimization, calibration, and pattern recognition).[23]

Most authors have found that the NN methods outperform not only MR but also PCA and PLS in the modeling of a large number of biochemical problems.[21,23] Therefore, they believed the NN models to be generally better than the MR models. This can be seen from the exponentially increasing number of papers in which the authors use NNs instead of MR and other methods.[23] While comparing NNs and MR, the authors often forget that they compare a highly nonlinear technique (NNs) with linear one (MR). In most cases, the authors have not attempted to introduce nonlinearities (only

† E-mail address: lucic@faust.irb.hr.
‡ E-mail address: trina@rudjer.irb.hr.

10.1021/ci980090f CCC: $18.00
Published on Web 11/19/1998

squares of initial descriptors were added but not their cross-products) and select the best possible MR model from the initial set of descriptors in order to see how good such a model is compared to those obtained by NNs. Therefore, the quality of the MR method was usually misjudged since the critical opinion was reached by considering models which were not the best possible MR models that could be obtained (except for very small sets of descriptors). We have shown in our previous reports[24−27] that by selecting the best possible descriptors to be used in MR modeling one obtains better models than those obtained using the usual approximate procedures for choosing descriptors.[28−31] Additionally, the selection of the best possible descriptors increases the stability of the coefficients in the MR model and thus the accuracy of the model also increases. In addition, all methods (NN, PLS, PCA) other than MR (except pure GA with a MR-like model) are not easy to relate from equation to equation, because relationships between the chemical structure and the activity of molecules are much more complex than in the case of MR (expressed by latent variables which vary from a model to model).[32]

In this report we will analyze the usefulness of the multivariate regression (MR) method in modeling the antileukemic activity of carboquinones against lymphoid leukemia in mice[33,34] and the tranquilizer activity of benzodiazepine derivatives.[34] The aim of this paper is (i) to present a new effective way for selecting the best possible descriptors in MR models, applicable in the cases when it is needed to select one or several models with six descriptors from a set of 100 descriptors; (ii) to present a new way for a stepwise selection of descriptors in the next-step model by addition of one, two, three, ... *i* new descriptors to all descriptors selected up to the previous-step model, where *i* is limited by the total number of descriptors and by the computer resources at one's disposal; (iii) to compare the performance of these MR models with other QSAR models such as earlier MR models obtained using approximate MR procedures,[33] models obtained using NN algorithms,[34−36] and the FUNCLINK method.[37]

METHODS

Data Sets. To show the usefulness of a new approach to building MR models, we decided to use the data well-studied both by NN and traditional MR methods and to compare the results. For this purpose, the carboquinone and the benzodiazepine data sets, the same as in ref 34, were chosen.

The carboquinone data set contains six descriptors, physicochemical parameters: the molecular refractivity constant (MR_1) and hydrophobicity constant (π_1) of substituents at the 1-position, substituent constants (*F* and *R*) which correspond to field and resonance electronic effects of substituents at the positions 1 and 2, and the total hydrophobicity $\pi_{1,2}$ ($\pi_1 + \pi_2$) and refractivity $MR_{1,2}$ ($MR_1 + MR_2$) used to estimate the steric effects of functional groups at the 1- and 2-positions. There are 39 different 2,5-bis(1-aziridinyl)-*p*-benzoquinones (position of the substituents can be seen from the chart shown in ref 34). Four kinds of biological data (log1/*c*) were used: the minimum effective dose (MED) and optimal dose (OD) on a chronic treatment schedule (chronic injection, ci) and those in single injection (si). There were 37 (MED-ci), 37 (OD-ci), 35 (MED-si), and 37 (OD-si) compounds with experimental biological data (antileukemic activity against lymphoid leukemia) in each data set.[33,34]

The benzodiazepine data sets contain seven physicochemical parameters (similar to those for carboquinones and denoted MR3, π_3, MR7, σ_3, *F*4, *R*4, and *I*1) which were estimated for 60 different benzodiazepines. These structural parameters and the biological data are the same as those in the literature.[34,38] The position in the structure of benzodiazepines (shown on the chart in ref 34) for which the parameters are calculated is indicated by the number after the sign of each parameter, as can be seen from the chart shown in ref 34. Activities are expressed as anti-pentylenetetrazole effect (anti-pent), anti-fighting effect (anti-fight), and clined screen test (clined-scr). These data sets contain measured activities for 57, 53, and 53 compounds, respectively.

In addition to six (for carboquinones) or seven (for benzodiazepines) structural parameters (descriptors), other authors used the squares of each initial descriptor as input in the NN. In this paper, the data sets are enlarged by the cross-product terms of 6 (or 12) and 7 (or 13) initial descriptors for carboquinone and benzodiazepine data sets, respectively. This will be discussed in detail in the section Results and Discussion. In contrast to NNs, there is no need to rescale inputs in the MR models, so all the given MR models can be checked outright.

Multivariate Regression. A regression equation is an equation for estimating a dependent variable A^{est} (which is an approximation to experimentally determined biological activities, *A*) from independent variables called descriptors (d_i, $i = 1, ..., I$). This relationship is expressed by a linear combination of descriptors:

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

The coefficients of contribution c_i of each single descriptor and their errors Δc_i are determined by the least-squares method. The model described by eq 1 is obtained by fitting experimental values of biological activity *A* to linear (or parabolic, or cubic, ...) functions of initial descriptors (physicochemical properties of substituent groups). The functions are parabolic or cubic or higher order curves if not only the initial descriptors are used but also their second-, third-, or higher order terms as well as cross-products. The cross-product terms correspond to interactions between physicochemical properties of substituent groups. The fitting quality is measured by the coefficient correlation (*R*) and standard error of estimate (S), since the quality of prediction for a new molecule from the same class is usually measured by a cross-validated correlation coefficient R_{cv} and 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. On the other hand, the cross-validated parameter R_{cv} will decrease and S_{cv} will increase. This indicates that if one wants to have a good model, one has to obtain a simple MR model, which will guarantee reliable prediction of the biological activity for a new molecule. 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 (for

significant descriptor $c_i \gg \Delta c_i$, usually, $c_i \cong 10 \cdot \Delta c_i$), where N can be very large, even 400.[10,11]

Selection of Descriptors for the Best MR Models. A computer program has been developed by which one can select the best MR model with I descriptors from the set of N descriptors. The number of such models with I descriptors is $N!/(N-I)!\cdot I!$. The best MR models are selected in the orthogonal basis, because in this basis the procedure is simpler and faster. To illustrate this point, we consider, as an example, MR modeling of a molecular activity A with two descriptors, d_1 and d_2.

The correlation coefficient (R) and the standard error of estimate (S) between the experimental activity A and estimated activity A^{est} calculated by MR from descriptors d_1 and d_2 are given as follows[25]

$$R(A, A^{est}) = [(R_{A,d_1}{}^2 + R_{A,d_2}{}^2 - 2R_{A,d_1}R_{A,d_2}R_{d_1,d_2})/(1 - R_{d_1,d_2}{}^2)]^{1/2} \quad (2)$$

$$S(A, A^{est}) = [M/(M-I-1)]^{1/2} \cdot S_A \cdot [(1 - R(A, A^{est})^2]^{1/2} \quad (3)$$

where M is the number of molecules in the set for which the model is derived, I denotes the number of descriptors used in the model (in the above case $I = 2$), and S_A is the standard deviation of A.

Because $R_{d_1,d_2} = 0$, eq 2 in the orthogonal basis reduces to much simpler expressions:

$$R(A, A^{est}) = [R_{A,d_1}{}^2 + R_{A,d_2}{}^2]^{1/2} \quad (4)$$

Simplification is evident even in this case of two descriptors, and it is considerable in the case of models with a large number of descriptors. The use of the orthogonal descriptor basis substantially facilitated the writing of the computer program *CROMRsel.f* (**CRO**atian **M**ulti**R**egression **sel**ection of descriptors) and also accelerated its execution. Thus, by the use of *CROMRsel.f* we can select the best model with six descriptors among 10^9 possible models (it takes about 10 h on Hewlett-Packard 9000/E55 computer, which is configured as a server) or the best model with five out of 104 descriptors ($\sim 10^8$ models, what takes 28 CPU min). Therefore, if we wish to express a certain physical or chemical property, or biological activity of a group of molecules as linear combination of descriptors, the problem we face is the selection of a set of I descriptors ($I = 1, ..., N$) from the set of N descriptors[11,26] which best approximate a given property or activity.[27,28] This problem was considered by a number of authors[6] but perhaps most consistently by Randić;[29] however, they gave no instructions (algorithm, computer program) how to solve any problem of real complexity (selection of descriptors in a large descriptor space).

Using the *CROMRsel.f* program in this paper the best possible models with 1, 2, ..., 6 descriptors were selected from a data sets containing 104 descriptors. After that, we have to decide which of these estimated models is most reliable for the application in predicting properties or activities of a new molecule. Namely, an increase of the number of descriptors in the model from I to $I + 1$ necessarily increases the value of R. This, however, does not mean a necessary decrease of the S-value, since in the computation of S (see eq 3) the number of descriptors used in the model is involved. Very often in QSAR reports, factor $[M/(M - I - 1)]^{1/2}$ is not taken into account in the evaluation of the S-value. This factor was not used by the authors of models with which we made the comparison and also in this paper either, to compare all the models on equal footing. Therefore, the decision on the number of descriptors to be used in the modeling must also rest on other criteria for checking the quality of models, such as the cross-validation technique (see discussion later in the text).

Stepwise Selection of *i* by *i* Descriptors for the MR Models. There are many ways of enhancing the capabilities of MR equations by taking into account higher order and cross-product terms, obtained by raising to a power or multiplying of initial descriptors. In our opinion, such terms are not used in practice because the enlargement of descriptor sets causes the problem of descriptor selection. Generally, the selection of a small number of important descriptors from a large initial set of descriptors is a permanent problem of modeling, which limits both the accuracy of models in the training procedure and their reliability in prediction.

The selection of novel descriptors for the MR model on the descriptors previously selected by the use of *CROMRsel.f* program was carried out not only in the *one by one* (as it is usually done in the stepwise MR procedure) but also in the *i by i* manner. The value of i is limited by the total number of descriptors N, by the number of descriptors I one wants to use in the MR model and by the computer resources at one's disposal. By the application of such a procedure, coded in the computer program *CROMRiisel.f* (**CRO**atian **M**ulti-**R**egression ***i*** by ***i*** **sel**ection of descriptors), one can obtain very good MR models which include all powers of second-, third-, fourth-, ... *n*th-order and cross-products of initial descriptors.

Gram−Schmidt Orthogonalization. We first carried out the orthogonalization of descriptors using the Gram−Schmidt algorithm for vectors, as initially described by Randić.[39] However, the orthogonalization of descriptors was described much more clearly by Szabo and Ostlund, and we followed their method.[40] Using their notation, vectors $|1\rangle$ and $|2\rangle$ can be orthogonalized as

$$|II\rangle = \frac{r_{12}|1\rangle - |2\rangle}{\sqrt{1 - r_{12}^2}} \quad (5)$$

where $|II\rangle$ is descriptor $|2\rangle$ orthogonalized against descriptor $|1\rangle$, while r_{12} is the correlation coefficient between the nonorthogonal descriptors $|1\rangle$ and $|2\rangle$. After this transformation descriptor $|1\rangle$ remains unchanged as well as the correlation coefficient between this descriptor and activity A, since it is the first descriptor against which the second descriptor was orthogonalized. The same procedure is repeated for all the remaining descriptors in the order in which we want them to be orthogonalized against the first descriptor.

The orthogonalization results in the change of the correlation coefficient between activity A and the orthogonalized descriptor $|II\rangle$. The new correlation coefficient R'_2 can be computed by

$$R'_2 = \langle A|II\rangle = \frac{r_{12}\langle A|1\rangle - \langle A|2\rangle}{\sqrt{1-r_{12}^2}} = \frac{r_{12}R_1 - R_2}{\sqrt{1-r_{12}^2}} \quad (6)$$

where $\langle A|$ is a symbolic representation of activity A as the bra vector, R_1 and R_2 stand for the correlation coefficients between property $\langle A|$ and nonorthogonal descriptors $|1\rangle$ and $|2\rangle$, while r_{12} is, like in eq 5, the correlation coefficient between the nonorthogonal descriptors $|1\rangle$ and $|2\rangle$. In the same manner, the correlation coefficients between activity A and the remaining descriptors orthogonalized against the first one must be recalculated.

The new correlation coefficients between all descriptors orthogonalized against the first descriptor can be computed only by the use of initial correlation coefficients (calculated between nonorthogonalized descriptors) in the following way. Let us denote the third descriptor in the orthogonalization order after orthogonalization against the first descriptor $|1\rangle$ as $|III_1\rangle$. When this descriptor is orthogonalized against descriptor $|II\rangle$, it will be denoted as $|III\rangle$. To carry out the orthogonalization, it is necessary to calculate the correlation coefficient r'_{23} between $|II\rangle$ and $|III_1\rangle$. This coefficient is defined by

$$r'_{23} = \langle III_1|II\rangle = \frac{-r_{12}r_{13} + r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}} \quad (7)$$

In the same manner we can recalculate the correlation coefficients between descriptor $|II\rangle$ and all the remaining descriptors that have been orthogonalized against the first descriptor. Then, each of the remaining descriptors is orthogonalized against descriptor $|II\rangle$ according to eq 5. This procedure repeats until all available descriptors are orthogonalized mutually in an a priori defined order. Finally, the correlation coefficients between orthogonalized descriptors and activity $\langle A|$ are used for the evaluation of the correlation coefficient of MR model using eq 4.

Neural Networks as a Nonlinear Technique. The NNs consisting solely of input and output layers are equivalent to multivariate regression.[34,41,42] To surpass the results obtained by linear operations, it is necessary to introduce a nonlinear operation into the NNs. Introduction of a hidden layer in an NN architecture causes transition from linear to nonlinear processing of input data. On the other hand, addition of the hidden layer increases the number of parameters (weights) which have to be optimized in the learning procedure. With too few hidden units, the network will not be able to extract the relevant nonlinear features, but with too many hidden units NN will contain too many weights which cause overfitting (but not generalizing) and NN "memorizes" individual data points. To conclude, we can say that the NN with a hidden layer is an intrinsically nonlinear technique. Comparison of such a method with MR is correct only if nonlinear transformations of initial descriptors are taken into account in the MR as well. This is done in the present report.

Three types of NN models were used in the comparison between NNs and MR:

(1) Four NN models for the carboquinone and three for the benzodiazepine data sets from the paper of Aoyama et al.[34] were obtained with the back-propagation algorithm and NN architecture containing one hidden layer. In the hidden layer, there were 12 or 26 neurons (two models) for the carboquinone and 28 neurons for the benzodiazepine data sets. The number of weights in an NN with bias neuron can be calculated by

$$P = (I+1)\cdot H + (H+1)\cdot O \quad (8)$$

where P, I, H, and O are the numbers of adjustable parameters, input, hidden, and output units, respectively.[41] According to eq 8, total numbers of weights in NNs from paper[34] were 96 or 364 for the carboquinone and 420 for the benzodiazepine data sets. From the comparison of these NN models with the MR models published by Yoshimoto et al.,[33] Aoyama et al.[34] concluded that the former are better.

(2) Four NN models for the carboquinone and three for the benzodiazepine data sets from the paper of Peterson[35] were obtained with counter propagation NNs (architectures of these NNs are different from those described in ref 34, see ref 35 for detail description). The author stated that 296 weights were optimized in the counter-propagation NNs.[35] From the comparison of these counter-propagation NN models with back-propagation NN[34] and MR models,[33] Peterson concluded that all have similar predictive capabilities.[35]

(3) One NN model for the carboquinone data set from the paper of Tetko et al.[36] was obtained with a feed-forward NN trained by the back-propagation algorithm. Besides, the authors used an algorithm for (only important) variable selection. Moreover, the final NN model was obtained as an average of 500 NNs. In each single NN there was one hidden layer with 10 neurons. The best NN model in the modeling of the chronic injection minimum effective dose (MED-ci) for the carboquinone data set was the model with four input descriptors (variables) and 61 optimized weights. From the comparison of these NN models[36] with the back-propagation NN models published by Aoyama et al.,[34] MR, NN, and FUNCLINK models published by Liu et al.,[37] Tetko et al.[36] concluded that (i) their NN models were better than other NN models (according to statistical parameters that the authors calculated), (ii) their NN models were especially better than MR models, and (iii) their NN models were slightly better or similar to FUNCLINK models.[37]

Statistical Indicators for the Quality of Models. The quality of the fit of the training data is expressed by the correlation coefficient, R, or standard error of estimate, S', computed as

$$S' = \sqrt{\frac{\sum_{i=1}^{M}(A_i - A_i^{\text{est}})^2}{M}} \quad (9)$$

where A_i, A_i^{est}, and M denote experimental activities, estimated activities, and the total number of cases (molecules) considered, respectively. However, a more important measure of the predictive quality are the cross-validated statistical parameters—correlation coefficient R_{cv} and standard error S'_{cv} (eq 10), which are calculated using experimental A_{cv_i} and

Table 1. Results for the Carboquinone MED Chronic Injection Data Set (**MED-ci**, 37 Compounds): Comparison of Several (the Best Possible) Multivariate Regression Models Obtained by the Program *CROMRsel.f* (Parts A−C) with Several NN, FUNCLINK, and "Old" Multivariate Regression Models (Part D)

A. The Best Possible MR Models Selected from the **CARB-12** Data Set: Six Initial Descriptors and Their Squares[b]

I	*R*	S'	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
5	0.9329	0.225	0.9026	0.278	d3, d4, d6, d8, d10
6	0.9438	0.207	0.9121	0.263	d3, d4, d6, d8, d10, d12
7	0.9462	0.202	0.9121	0.265	d3, d9, d11, d4, d6, d8, d10
8	0.9465	0.202	0.9124	0.264	d3, d5, d9, d11, d4, d6, d8, d10

B. The Best Possible MR Models Selected from the **CARB-27** Data Set: Six Initial Descriptors and Their Cross-Product Terms[b]

I	*R*	S'	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
5	0.9476	0.200	0.9317	0.229	d3, d1d7, d3d7, d3d11, d9d9
6	0.9508	0.194	0.9382	0.218	d3, d7, d3d7, d7d7, d7d11, d9d9
7	0.9590	0.177	0.9441	0.208	d3, d1d1, d1d7, d3d11, d5d7, d7d7, d9d9

MED-ci = (6.587 ± 0.088) + (−0.937 ± 0.094)**d3** + (−0.074 ± 0.021)**d1d1** + (0.62 ± 0.17)**d1d7** + (−1.01 ± 0.27)**d3d11** + (0.353 ± 0.079)**d5d7** + (−1.08 ± 0.28)**d7d7** + (−3.33 ± 0.82)**d9d9**

C. The Best Possible MR Models Selected from the **CARB-90** Data Set: 12 Descriptors from the CARB-12 Data Set and Their Cross-Product Terms[b]

I	*R*	S'	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
3	0.9217	0.243	0.8920	0.284	d3, d4d12, d8d12
4	0.9462	0.202	0.9291	0.231	d3, d3d11, d4d8, d7d11
4	0.9487	0.198	0.9300	0.230	d3, d3d11, d4d8, d8d12
5	0.9548	0.186	0.9444	0.206	d3, d3d11, d4d8, d8d11, d9d11
5	0.9541	0.187	0.9459	0.203	d3, d2d2, d3d12, d4d4, d9d12
6	0.9635	0.167	0.9550	0.186	d1d10, d2d2, d2d11, d3d8, d5d7, d5d8
6	0.9638	0.167	0.9553	0.185	d3, d12, d1d2, d2d12, d4d7, d7d11
6	0.9644	0.165	0.9561	0.183	d3, d1d2, d2d12, d4d8, d7d11, d11d11

MED-ci = (6.766 ± 0.072) + (−0.584 ± 0.040)**d3** + (−0.0079 ± 0.0011)**d1d2** + (0.642 ± 0.093)**d2d12** + (0.0099 ± 0.0017)**d4d8** + (1.11 ± 0.24)**d7d11** + (−1.87 ± 0.34)**d11d11**

D. NN, FUNCLINK, and "old" MR Models

	R	S'	R_{cv}	S'_{cv}
NN1[c]		0.21		0.26
NN2[d]		0.21		
NN3[e]		0.08		0.33
NN4[f]		0.06		0.24[g]
NN5[h]	0.98		0.93	
NN6[i]	0.98		0.92	
NN7[j]	0.94		0.86	
FUNCLINK[k]	0.95		0.94	
MLR1[l]	0.91		0.80	
MLR2[l]	0.89		0.84	
MLR3[m]		0.24		0.25

[a] *I* = the number of descriptors in multivariate regression (MR) models; *R* = correlation coefficient; R_{cv} = leave-one-out (cross-validated) correlation coefficient; S' and S'_{cv} = standard error (root-mean-square error) and leave-one-out standard error between known and (by model) predicted data were calculated using eqs 9 and 10. Description for MED-ci is given in text (see sections Data Sets or Results and Discussion). [b] Descriptors are denoted as follows: $d1 = MR_{1,2}$; $d2 = (MR_{1,2})^2$; $d3 = \pi_{1,2}$; $d4 = (\pi_{1,2})^2$; $d5 = \pi_2$; $d6 = (\pi_2)^2$; $d7 = MR_1$; $d8 = (MR_1)^2$; $d9 = F$; $d10 = F^2$; $d11 = R$; $d12 = R^2$. [c] Result by the NN with six input parameters (set A) from ref 34. [d] Result by the NN with six physicochemical parameters and their squares as input (set B) from ref 34. [e] Result by the NN with six physicochemical input parameters from ref 35. [f] Result by the NN with six physicochemical parameters and their squares as input from ref 35. [g] The best leave-one-out (cross-validated) result obtained by the NN from ref 35. [h] The best result by the NN with four selected physicochemical parameters as input from ref 36. [i] Result by the NN with five selected physicochemical parameters as input from ref 36. [j] Result by the NN (generalized delta rule net) with 10 hidden neurons from ref 37. [k] Result by the FUNCLINK (functional link net) from ref 37. [l] Results by the MLR were taken from refs 37 and 36. [m] Results by the MLR from refs 33, 34, and 35.

estimated A_{cvi}^{est} activities based on the leave-one-out cross-validation procedure

$$S'_{cv} = \sqrt{\frac{\sum_{i=1}^{M}(Acv_i - Acv_i^{est})^2}{M}} \qquad (10)$$

It should be noted that the factor which takes into account the number of descriptors in the models (see eq 3) is not used in either eq 9 or eq 10.

Implementation. Computer programs for the selection of the best possible descriptors in MR models by the use of orthogonalization procedure *CROMRsel.f* (**CRO**atian **M**ulti**R**egression **sel**ection of descriptors) as well as for the new (*i by i*) stepwise selection of descriptors in MR models *CROMRiisel.f* (**CRO**atian **M**ulti**R**egression ***i*** by ***i*** **sel**ection of descriptors) have been written in the FORTRAN **77**. All

Table 2. Results for the Carboquinone OD Chronic Injection Data Set (**OD-ci**, 37 Compounds): Comparison of Several (the Best Possible) Multivariate Regression Models Obtained by the Program *CROMRsel.f* (Parts A−C) with Several NN and "Old" Multivariate Regression Models (Part D)

A. The Best Possible MR Models Selected from the **CARB-12** Data Set: Six Initial Descriptors and Their Squares[b]

I	*R*	*S*′	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
6	0.9398	0.174	0.9038	0.223	d3, d7, d9, d11, d4, d6
7	0.9399	0.174	0.8949	0.233	d1, d3, d7, d9, d11, d4, d6
8	0.9400	0.174	0.8874	0.242	d1, d3, d7, d9, d11, d2, d4, d6

B. The Best Possible MR Models Selected from the **CARB-27** Data Set: Six Initial Descriptors and Their Cross-Product Terms[b]

I	*R*	*S*′	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
5	0.9408	0.173	0.9212	0.200	d3, d7, d9, d11, d3d7
6	0.9441	0.168	0.9246	0.197	d3, d7, d9, d11, d3d7, d5d5
7	0.9554	0.151	0.9324	0.185	d1, d3, d7, d9, d1d3, d5d5, d5d11

OD-ci = (6.58 ± 0.14) + (−0.199 ± 0.049)**d1** + (−1.04 ± 0.12)**d3** + (−0.295 ± 0.076)**d7** + (−1.57 ± 0.31)**d9** + (0.201 ± 0.029)**d1d3** + (−0.261 ± 0.042)**d5d5** + (−2.09 ± 0.42)**d5d11**

C. The Best Possible MR Models Selected from the **CARB-90** Data Set: 12 Descriptors from the CARB-12 Data Set and Their Cross-Product Terms[b]

I	*R*	*S*′	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
3	0.8916	0.231	0.8727	0.250	d2, d3d3, d3d4
4	0.9349	0.181	0.9243	0.195	d4, d3d8, d4d8, d5d12
5	0.9641	0.136	0.9522	0.156	d9, d1d11, d2d2, d3d7, d3d8
6	0.9788	0.105	0.9709	0.123	d1d2, d1d12, d3d4, d3d7, d4d8, d12d12
6	0.9797	0.102	0.9734	0.117	d1d12, d2d2, d3d4, d3d7, d4d8, d12d12

OD-ci = (5.506 ± 0.036) + (2.59 ± 0.21)**d1d12** + (−0.000725 ± 0.00009)**d2d2** + (−0.0223 ± 0.0023)**d3d4** + (−0.404 ± 0.028)**d3d7** + (0.0424 ± 0.0026)**d4d8** + (−4.24 ± 0.37)**d12d12**

D. NN and "old" MR Models

	R	*S*′	R_{cv}	S'_{cv}
NN1[c]		0.19		0.26
NN2[d]		0.18		
NN3[e]		0.02		0.27
NN4[f]		0.05		0.23[g]
MLR[h]		0.23		0.23

[a] See footnote in Table 1 for description of *I*, *R*, R_{cv}, *S*′, and S'_{cv}. Description for OD-ci is given in text (see sections Data Sets or Results and Discussion). [b] Descriptors are denoted as follows: $d1 = MR_{1,2}$; $d2 = (MR_{1,2})^2$; $d3 = \pi_{1,2}$; $d4 = (\pi_{1,2})^2$; $d5 = \pi_2$; $d6 = (\pi_2)^2$; $d7 = MR_1$; $d8 = (MR_1)^2$; $d9 = F$; $d10 = F^2$; $d11 = R$; $d12 = R^2$. [c] Result by the NN with six input parameters (set A) from ref 34. [d] Result by the NN with six physicochemical parameters and their squares as input (set B) from ref 34. [e] Result by the NN with six physicochemical input parameters from ref 36. [f] Result by the NN with six physicochemical parameters and their squares as input from ref 36. [g] The best leave-one-out (cross-validated) result obtained by the NN from ref 36. [h] Results by the MLR from refs 33, 34, and 36.

calculations were done on a Hewlett-Packard 9000/E55 server.

RESULTS AND DISCUSSION

New MR models will be presented, and a comparison with several NN models will be given considering two sets of compounds: (1) four sets of models for the carboquinone biological activity and (2) three sets of models for the benzodiazepine biological activity. All the models for carboquinones were obtained by selecting the best possible descriptors, using the *CROMRsel.f* program. The models for benzodiazepines were obtained by the stepwise *i by i* selection of descriptors, using the *CROMRiisel.f* program.

A. QSAR in the Carboquinone Data Set. The QSAR models for the four biological activities (MED-ci, OD-ci, MED-si, and OD-si) in the carboquinone data set are obtained by selection of absolutely the best descriptors from the three data sets of descriptors: (**A**) six initial descriptors and their squares (12-descriptor set, **CARB-12**, Tables 1A−4A); (**B**) six initial descriptors and their cross-product terms (27-descriptor set, **CARB-27**, Tables 1B−4B); (**C**) 12 descriptors from a 12-descriptor set and their cross-product terms (90-descriptor set, **CARB-90**, Tables 1C−4C).

MED-ci Models. Multiregression models of the 37 carboquinones for the minimum effective dose chronic injection (MED-ci) biological data obtained on 12-, 27-, and 90-descriptor data sets are shown in Table 1A−C. Statistical parameters of the other models (traditional MR, NN and FUNCLINK) are given in Table 1D. We did not select only one (the best possible) MED-ci model for each data set but also several almost equally good models, ranked by their correlation coefficients.

OD-ci models of the 37 carboquinones obtained in this work, and the results of NN and "old" MR, are shown in Table 2A−D. Like in the case of MED-ci models, we have presented several, almost equally good, nonlinear MR models.

The best possible **MED-si** and **OD-si** MR models of 35 and 37 carboquinones obtained for each of the three data sets as well as the NN and "old" MR models are given in Tables 3 and 4, respectively.

Table 3. Results for the Carboquinone MED Single Injection Data Set (**MED-si**, 35 Compounds): Comparison of Several (the Best Possible) Multivariate Regression Models Obtained by the Program *CROMRsel.f* (Parts A−C) with Several NN and "Old" Multivariate Regression Models (Part D)

A. The Best Possible MR Model Selected from the **CARB-12** Data Set: Six Initial Descriptors and Their Squares[b]

I	R	S'	R_{cv}	S'_{cv}	descriptors selected in MR model[b]
7	0.9297	0.237	0.8271	0.418	d1, d3, d4, d6, d8, d9, d12

B. The Best Possible MR Model Selected from the **CARB-27** Data Set: Six Initial Descriptors and Their Cross-Product Terms[b]

I	R	S'	R_{cv}	S'_{cv}	descriptors selected in MR model[b]
8	0.9450	0.210	0.9213	0.252	d1, d3, d1d9, d3d7, d3d9, d7d9, d7d11, d11d11

MED-si = (6.14 ± 0.20) + (−0.294 ± 0.073)**d1** + (−0.32 ± 0.13)**d3** + (−5.30 ± 0.85)**d1d9** + (0.212 ± 0.061)**d3d7** + (6.0 ± 1.6)**d3d9** + (5.9 ± 1.8)**d7d9** + (0.42 ± 0.40)**d7d11** + (1.453 ± 0.279)**d11d11**

C. The Best Possible MR Model Selected from the **CARB-90** Data Set: 12 Descriptors from the CARB-12 Data Set and Their Cross-Product Terms[b]

I	R	S'	R_{cv}	S'_{cv}	descriptors selected in MR model[b]
6	0.9474	0.206	0.9395	0.221	d9, d3d8, d5d7, d5d8, d9d12, d11d12

MED-si = (5.377 ± 0.062) + (−5.25 ± 0.67)**d9** + (−0.245 ± 0.046)**d3d8** + (−1.36 ± 0.13)**d5d7** + (0.93 ± 0.13)**d5d8** + (1.33 ± 0.90)**d9d12** + (−1.34 ± 0.21)**d11d12**

D. NN and "old" MR Models

	R	S'	R_{cv}	S'_{cv}
NN1[c]		0.25		0.48
NN2[d]		0.24		
NN3[e]		0.04		0.44
NN4[f]		0.20		0.38[g]
MLR[h]		0.27		0.26

[a] See footnote in Table 1 for description of I, R, R_{cv}, S', and S'_{cv}. Description for MED-si is given in text (see sections Data Sets or Results and Discussion). [b] Descriptors are denoted as follows: $d1 = MR_{1,2}$; $d2 = (MR_{1,2})^2$; $d3 = \pi_{1,2}$; $d4 = (\pi_{1,2})^2$; $d5 = \pi_2$; $d6 = (\pi_2)^2$; $d7 = MR_1$; $d8 = (MR_1)^2$; $d9 = F$; $d10 = F^2$; $d11 = R$; $d12 = R^2$. [c] Result by the NN with six input parameters (set A) from ref 34. [d] Result by the NN with six physicochemical parameters and their squares as input (set B) from ref 34. [e] Result by the NN with six physicochemical input parameters from ref 36. [f] Result by the NN with six physicochemical parameters and their squares as input from ref 36. [g] The best leave-one-out (cross-validated) result obtained by the NN from ref 36. [h] Results by the MLR from refs 33, 34, and 36.

Internal comparison of the MR models in parts A, B, and C of Tables 1−4 clearly shows that an efficient selection of descriptors over a larger data set of descriptors leads to a better final model. On the other hand, introduction of nonlinearities into the descriptor data set used for the selection of MR models is necessary if one wants to perform a meaningful comparison between MR and NN with a hidden layer (which is an intrinsically nonlinear technique). Nonlinearities are taken into account through calculation of cross-product terms (by simple multiplications of the initial descriptors relating to the same molecule).

From the comparison of the MR models (Tables 1−4, parts A−C) with the corresponding NN models (Tables 1D−4D), it is clearly seen that the former are much better. In addition, the MR models are much simpler. This enables a direct inspection of the significance of every descriptor in the model. The best possible MR models use only six nonlinear descriptors, with only seven adjustable parameters. From Tables 1 and 2, we can see that not only the best one but also several others MR models (second-, third- fourth-, ..., the best model) are superior to each of the NN, FUNCLINK, or "old" MR models.

It is interesting to see whether several initial descriptors (physicochemical parameters) are important or not for modeling the carboquinone biological activity. To test the possibilities of NN for variable selection, one of the NN models used the carboquinone data set and MED-ci biological activity.[36] Tetko et al.[36] have found that the $MR_{1,2}$ and π_2 are irrelevant parameters for the NN model of MED-ci.

From Table 1C, we can see that the most frequently used descriptors in the modeling of MED-ci are $d3 = \pi_{1,2}$ (which appears in almost all models), $d1 = MR_{1,2}$ or $d2 = (MR_{1,2})^2$, and $d11 = R$ or $d12 = R^2$. On the other hand, the most irrelevant are descriptors $d5 = \pi_2$, $d6 = (\pi_2)^2$, $d9 = F$, and $d10 = F^2$. This indicates that if one descriptor is irrelevant for NN, it does not mean that the same descriptor is irrelevant for the MR method. However, the ensemble of very good MR models from Table 1, which are superior to all other models, favors the MR selection of important descriptors.

The best OD-ci models selected from the CARB-90 data set (Table 2C) do not contain descriptors $d9 = F$, and $d10 = F^2$ (substituent constants which correspond to field electronic effects of substituents 1 and 2). The most frequent are descriptors $d3d7 = \pi_{1,2}MR_1$, $d12d12 = R^2R^2 = R$,[4] and $d1d12 = MR_{1,2}R^2$. We can see that the most important descriptors are cross-products of total hydrophobicity and the refractivity constant of substituent 1. Since all the values of MR_1 are of a positive sign as well as almost all the values of $\pi_{1,2}$ (except for five molecules, see ref 34), almost all the values of descriptor $d3d7$ are of a positive sign. Because the multiregression coefficients at descriptor $d3d7$ are of a negative sign, we conclude that this term will contribute to

Table 4. Results for the Carboquinone OD Single Injection Data Set (OD-si, 37 Compounds): Comparison of Several (the Best Possible) Multivariate Regression Models Obtained by the Program *CROMRsel.f* (Parts A−C) with Several NN and "Old" Multivariate Regression Models (Part D)

A. The Best Possible MR Model Selected from the **CARB-12** Data Set: Six Initial Descriptors and Their Squares[b]

I	R	S'	R_{cv}	S'_{cv}	descriptors selected in MR model[b]
7	0.9409	0.201	0.7990	0.382	d3, d4, d6, d7, d10, d11, d12

B. The Best Possible MR Model Selected from the **CARB-27** Data Set: Six Initial Descriptors and Their Cross-Product Terms[b]

I	R	S'	R_{cv}	S'_{cv}	descriptors selected in MR model[b]
7	0.9578	0.170	0.9458	0.193	d3, d9, d11, d3d3, d5d5, d7d9, d7d11

OD-si = (4.782 ± 0.076) + (−0.503 ± 0.084)**d3** + (−5.34 ± 0.67)**d9** + (−1.70 ± 0.19)**d11** + (0.066 ± 0.020)**d3d3** + (−0.095 ± 0.032)**d5d5** + (1.97 ± 0.58)**d7d9** + (0.37 ± 0.24)**d7d11**

C. The Best Possible MR Model Selected from the **CARB-90** Data Set: 12 Descriptors from the CARB-12 Data Set and Their Cross-Product Terms[b]

I	R	S'	R_{cv}	S'_{cv}	descriptors selected in MR model[b]
6	0.9564	0.173	0.9481	0.189	d3, d9, d11, d2d5, d4d4, d8d9

OD-si = (4.750 ± 0.072) + (−0.446 ± 0.052)**d3** + (−4.42 ± 0.39)**d9** + (−1.54 ± 0.16) **d11** + (−0.0030 ± 0.0014)**d2d5** + (0.00188 ± 0.00055)**d4d4** + (0.83 ± 0.17)**d8d9**

D. NN, and "old" MR Models

	R	S'	R_{cv}	S'_{cv}
NN1[c]		0.20		0.30
NN2[d]		0.17		
NN3[e]		0.08		0.47
NN4[f]		0.05		0.32
MLR[g]		0.24		0.23

[a] See footnote in Table 1 for description of I, R, R_{cv}, S', and S'_{cv}. Description for OD-si is given in text (see sections Data Sets or Results and Discussion). [b] Descriptors are denoted as follows: $d1 = MR_{1,2}$; $d2 = (MR_{1,2})^2$; $d3 = \pi_{1,2}$; $d4 = (\pi_{1,2})^2$; $d5 = \pi_2$; $d6 = (\pi_2)^2$; $d7 = MR_1$; $d8 = (MR_1)^2$; $d9 = F$; $d10 = F^2$; $d11 = R$; $d12 = R^2$. [c] Result by the NN with six input parameters (set A) from ref 34. [d] Result by the NN with six physicochemical parameters and their squares as input (set B) from ref 34. [e] Result by the NN with six physicochemical input parameters from ref 36. [f] Result by the NN with six physicochemical parameters and their squares as input from ref 36. [g] Results by the MLR from refs 33, 34, and 36.

a significant decrease of the OD-ci value for the molecule with high values of both $\pi_{1,2}$ and MR_1 descriptors. Furthermore, since descriptor $d12d12 = R^4$ has all the values of a positive sign and the corresponding multiregression coefficient of a negative sign, this term decreases the values of OD-ci for all the molecules. Contrary to the two above-mentioned significant descriptors, descriptor $d1d12 = MR_{1,2}R^2$ increases the values of OD-ci because $MR_{1,2}$, R^2 as well as the corresponding multiregression coefficient are of a positive sign.

Based on the comparison between the NN and multivariate linear regression models, the authors of the NN papers concluded that the carboquinone data set is quite linear because these two methods gave (more or less) similar results. However, from the results for the carboquinones presented in Tables 1−4 it is evident that (i) all the carboquinone data sets should be treated as nonlinear sets; (ii) it is possible to make an effective selection of a small number of significant descriptors from a large initial data set; (iii) introduction of nonlinear terms significantly improves the accuracy of MR models; and (iv) there are not only one but several MR models which are better than each of the published NN model.

B. QSAR in Benzodiazepine Data Sets. The QSAR models for the three biological activities (anti-pent, anti-fight, clined-scr, see Data Sets section or ref 34 for description) in the benzodiazepine data set were obtained by the application of the stepwise *i by i* selection of descriptors in MR models. For that purpose, the computer program *CROMRii-sel.f*, previously developed and described above (see Methods), was used. The initial data set contained seven descriptors (see Data Sets section). By introduction of nonlinear terms three data sets were created and used for the selection of the models: (**A**) 13-descriptor set (BENZO-13) which contains seven initial descriptors and the squares of all descriptors except of I-1 (namely, $(I\text{-}1)^2 = I\text{-}1$ because all the values of descriptor I-1 are 0 or 1, see ref 34); (**B**) 35-descriptor set (BENZO-35) which contains seven initial descriptors and their cross-product terms; (**C**) 104-descriptor set (BENZO-104) which contains 13 descriptors from the BENZO-13 and their cross-product terms.

Anti-Pent Models. Multiregression models of the 57 benzodiazepines for anti-pentylenetetrazole biological data obtained on 13-, 35-, and 104-descriptor data sets are shown in Table 5A−C. Statistical parameters of the other models (traditional MR and NN) are shown in Table 5D.

The best cross-validated MR model selected from the BENZO-13 data set is the model with only three descriptors (Table 5A). In addition, statistical parameters for the best possible MR models with six and with all 13 descriptors, selected from BENZO-13 data set, are given. The comparison of these MR models with NN models from Table 5D shows that latter are better. We assumed that in this case the NN algorithm found significant nonlinear relations between the input descriptors and included them in the NN anti-pent

Table 5. Results for the Benzodiazepine Anti-Pentylenetetrazole Data Set (Anti-Pent, 57 Compounds): Comparison of Several Multivariate Regression Models Obtained by the Program *CROMRsel.f* and by the Program *CROMRiisel.f* (Parts A−C) with Several NN and "Old" Multivariate Regression Models (Part D)

A. The Best Possible MR Models Selected from the BENZO-13 Data Set: Seven Initial Descriptors and Their Squares[b]

I	*R*	S'	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
3	0.6864	0.574	0.6345	0.614	d4, d7, d9
6	0.7461	0.526	0.6800	0.582	d2, d3, d8, d9, d11, d12
13	0.8150	0.458	0.4070	1.121	model with all 13 descriptors

B. The Best Possible MR Models Selected from the BENZO-35 Data Set: Seven Initial Descriptors and Their Cross-Product Terms[b]

I	*R*	S'	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
5	0.7907	0.484	0.7487	0.526	d1d9, d1d13, d3d7, d5d7, d7d7
6	0.8064	0.468	0.7579	0.519	d1d9, d1d11, d1d13, d3d7, d5d7, d7d7
7	0.8379	0.432	0.7854	0.493	d1, d3, d1d9, d1d11, d5d7, d7d7, d11d11

anti-pent = (4.92 ± 0.22) + (−0.76 ± 0.16)**d1** + (0.40 ± 0.15)**d3** + (1.91 ± 0.53)**d1d9** + (−4.0 ± 1.0)**d1d11** + (−0.46 ± 0.15)**d5d7** + (3.10 ± 0.57)**d7d7** + (−7.4 ± 1.9)**d11d11**

C. The Best Possible MR Models with 4−6 Descriptors Selected by *CROMRsel.f* and the MR Models with 11, 15 and 19 Descriptors Obtained by Stepwise (*i* by *i*) Algorithm (Coded in *CROMRiisel.f*) from the BENZO-104 Data Set: 13 Descriptors from the BENZO-13 Data Set and Their Cross-Product Terms[b]

I	*R*	S'	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
4	0.7891	0.486	0.7592	0.516	d1d4, d1d7, d2d7, d2d9
5	0.8116	0.462	0.7769	0.499	d8, d1d11, d3d8, d6d8, d12d12
6	0.8504	0.416	0.8161	0.460	d1d8, d1d11, d2d8, d6d8, d7d11, d12d12
11	0.9043	0.337	0.8630	0.404	d9, d13, d1d8, d1d11, d2d8, d4d13, d5d10, d6d8, d7d8, d7d11, d12d12
15	0.9263	0.298	0.8863	0.371	d5, d9, d13, d1d8, d1d11, d2d8, d4d13, d5d5, d5d6, d5d10, d6d6, d6d8, d7d8, d7d11, d12d12

anti-pent = (8.3 ± 1.7) + (−13.1 ± 5.2)**d5** + (1.39 ± 0.35)**d9** + (−3.4 ± 1.3)**d13** + (29.6 ± 5.7)**d1d8** + (−11.0 ± 2.0)**d1d11** + (−19.7 ± 3.1)**d2d8** + (0.38 ± 0.12)**d4d13** + (12.9 ± 4.8)**d5d5** + (−4.9 ± 1.7)**d5d6** + (−1.37 ± 0.65)**d5d10** + (0.63 ± 0.21)**d6d6** + (−0.43 ± 0.14)**d6d8** + (−11.0 ± 3.6)**d7d8** + (9.0 ± 2.2)**d7d11** + (−33.1 ± 6.6)**d12d12**

19	0.9415	0.266	0.8437	0.462	d5, d9, d13, d1d8, d1d11, d2d7, d2d8, d4d9, d4d13, d5d5, d5d6, d5d10, d6d6, d6d8, d7d8, d7d11, d9d9, d9d10, d12d12

D. NN and "old" MR Models

	R	S'	R_{cv}	S'_{cv}
NN1[c]		0.33		0.53
NN2[d]		0.16		0.59
NN3[e]				0.53
MLR[f]		0.61		0.47

[a] I = the number of descriptors in multivariate regression (MR) models; R = correlation coefficient; R_{cv} = leave-one-out (cross-validated) correlation coefficient; S' and S'_{cv} = standard error (root-mean-square error) and leave-one-out standard error between known and (by model) predicted data were calculated using eqs 9 and 10. Description for anti-pent is given in text (see sections Data Sets or Results and Discussion). [b] Descriptors are denoted as follows: $d1 = MR\text{-}3$; $d2 = (MR\text{-}3)^2$; $d3 = \pi - 3$; $d4 = (\pi - 3)^2$; $d5 = MR\text{-}7$; $d6 = (MR\text{-}7)^2$; $d7 = \sigma\text{-}3$; $d8 = (\sigma\text{-}3)^2$; $d9 = F\text{-}4$; $d10 = (F\text{-}4)^2$; $d11 = R\text{-}4$; $d12 = (R\text{-}4)^2$; $d13 = I\text{-}1$; note that $(I\text{-}1)^2 = I\text{-}1$ because all the values of descriptors I-1 are 0 or 1 (see ref 34). [c] Result by the back-propagation NN with seven physicochemical parameters and their squares as input from ref 34. [d] Result by the counter-propagation NN with seven physicochemical parameters and their squares as input from ref 36. [e] The best leave-one-out (cross-validated) result obtained by the NN with 32 neurons in the hidden layer from ref 36. [f] Results by the MLR from refs 34 and 36.

model. Therefore, to improve the quality of MR models, we enlarged the initial data set containing 13 descriptors by inclusion of nonlinear (cross-product) terms, which gave us two new data sets: BENZO-35 and BENZO-104.

For the BENZO-35 data set, the best possible MR models up to seven descriptors are selected, and only the best three models with five, six, and seven descriptors are shown in Table 5B. It is clear that these models have lower values of the cross-validated standard error S'_{cv} than the NN models. The value of S'_{cv} for the MLR model in ref 35 is incorrect. Namely, for this model, the author has given lower values for S'_{cv} than for the non-cross-validated standard error S', what is impossible (Table 5D). This model corresponds to the MR models that are given in Table 5A, and because of that it seems that the values of S'_{cv} should be > 0.70 (perhaps 0.74).

For the BENZO-104 data set, we first obtained the best possible MR models up to six descriptors by the *CROMRsel.f* program. Then, the computer program *CROMRiisel.f* for the stepwise selection of descriptors in the *i by i* manner was applied starting from the best possible MR model with six descriptors. The values of *i* in the *i by i* stepwise selection of descriptors were chosen to be 5, 4, and 4 for the first, second, and third steps, respectively. So, through three steps by addition of 5, 4, and 4 new descriptors in each step we obtained the MR models with 11, 15, and 19 descriptors

Table 6. Results for the Benzodiazepine Anti-Fighting Effect Data Set (Anti-Fight, 53 Compounds): Comparison of Several Multivariate Regression Models Obtained by the Program *CROMRsel.f* and by the Program *CROMRiisel.f* (Part A) with Several NN and "Old" Multivariate Regression Models (Part B)

A. The Best Possible MR Models with Five and Six Descriptors Selected by *CROMRsel.f* and the MR Models with 11, 15, and 19 Descriptors Obtained by the Stepwise (*i by i*) Algorithm (Coded in *CROMRiisel.f*) from the BENZO-104 Data Set: 13 Descriptors from BENZO-13 Data Set and Their Cross-Product Terms[b]

I	R	S'	R_{cv}	S'_{cv}	descriptors selected in MR models[b]
5	0.8225	0.316	0.7851	0.345	d4, d5d7, d7d13, d7d8, d10d11
6	0.8368	0.304	0.7877	0.344	d4, d5d7, d7d13, d7d8, d10d11, d4d10
		anti-fight $= (3.58 \pm 0.13) + (1.30 \pm 0.26)$**d9** $+ (-0.81 \pm 0.13)$**d3d4** $+$ (-1.34 ± 0.27)**d4d7** $+ (4.98 \pm 0.53)$**d4d11** $+ (-6.7 \pm 2.0)$**d6d12** $+ (3.2 \pm 1.7)$**d9d12**			
11	0.9079	0.233	0.8617	0.287	d1d11, d1d10, d2d12, d4, d4d10, d5d7, d7d13, d7d8, d9d12, d10d10, d10d11
15	0.9195	0.219	0.8338	0.315	d1d11, d1d10, d2d12, d3d13, d4, d4d4, d4d10, d5d7, d5d10, d7d8, d7d13, d9d12, d10d10, d10d11, d11d6
19	0.9620	0.152	0.8099	0.355	d1d10, d1d11, d2d12, d3d8, d3d13, d4, d4d4, d4d10, d5d7, d5d10, d5d11, d6d10, d7d8, d7d12, d7d13, d9d12, d10d10, d10d11, d11d6

B. NN and "old" MR Models

	R	S'	R_{cv}	S'_{cv}
NN1[c]		0.21		0.38
NN2[d]		0.15		0.47
NN3[e]				0.44
MLR[f]		0.39		0.40

[a] See footnote in Table 5 for description of I, R, R_{cv}, S', and S'_{cv}. Description for anti-fight effect is given in text (see sections Data Sets or Results and Discussion). [b] Descriptors are denoted as follows: $d1 = MR$-3; $d2 = (MR\text{-}3)^2$; $d3 = \pi - 3$; $d4 = (\pi - 3)^2$; $d5 = MR$-7; $d6 = (MR\text{-}7)^2$; $d7 = \sigma$-3; $d8 = (\sigma\text{-}3)^2$; $d9 = F$-4; $d10 = (F\text{-}4)^2$; $d11 = R$-4; $d12 = (R\text{-}4)^2$; $d13 = I$-1; note that $(I\text{-}1)^2 = I$-1 because all the values of descriptors I-1 are 0 or 1 (see ref 34). [c] Result by the back-propagation NN with seven physicochemical parameters and their squares as input from ref 34. [d] Result by the counter-propagation NN with seven physicochemical parameters and their squares as input from ref 36. [e] The best leave-one-out (cross-validated) result obtained by the NN with 32 neurons in the hidden layer from ref 36. [f] Results by the MLR from refs 34 and 36.

Table 7. Results for the Benzodiazepine Clined Screen Test Effect Data Set (Clined-scr, 53 Compounds): Comparison of the Multivariate Regression Model Obtained by the Program *CROMRsel.f* with Several NN and "Old" Multivariate Regression Models (Part B)

A. The Best Possible MR Model with Six Descriptors Selected by *CROMRsel.f* from the BENZO-104 Data Set: 13 Descriptors from the BENZO-13 Data Set and Their Cross-Product Terms[b]

I	R	S'	R_{cv}	S'_{cv}	descriptors selected in MR model[b]
6	0. 8532	0.302	0. 8126	0. 338	d4, d3d4, d5d7, d8d8, d8d11, d12d13
		clined-scr $= (3.06 \pm 0.11) + (0.96 \pm 0.17)$**d4** $+ (0.66 \pm 0.10)$**d3d4** $+$ (-0.79 ± 0.11)**d5d7** $+ (5.54 \pm 0.62)$**d8d8** $+ (-5.7 \pm 1.1)$**d8d11** $+ (-5.0 \pm 1.2)$**d12d13**			

B. NN and "old" MR Models

	R	S'	R_{cv}	S'_{cv}
NN1[c]		0.31		0.41
NN2[d]		0.15		0.52
NN3[e]				0.43
MLR[f]		0.41		0.41

[a] See footnote in Table 5 for description of I, R, R_{cv}, S', and S'_{cv}. Description for clined-scr test effect is given in text (see sections Data Sets or Results and Discussion). [b] Descriptors are denoted as follows: $d1 = MR$-3; $d2 = (MR\text{-}3)^2$; $d3 = \pi - 3$; $d4 = (\pi - 3)^2$; $d5 = MR$-7; $d6 = (MR\text{-}7)^2$; $d7 = \sigma$-3; $d8 = (\sigma\text{-}3)^2$; $d9 = F$-4; $d10 = (F\text{-}4)^2$; $d11 = R$-4; $d12 = (R\text{-}4)^2$; $d13 = I$-1; note that $(I\text{-}1)^2 = I$-1 because all the values of descriptors I-1 are 0 or 1 (see ref 34). [c] Result by the back-propagation NN with seven physicochemical parameters and their squares as input from ref 34. [d] Result by the counter-propagation NN with seven physicochemical parameters and their squares as input from ref 36. [e] The best leave-one-out (cross-validated) result obtained by the NN with 32 neurons in the hidden layer from ref 36. [f] Results by the MLR from refs 34 and 36.

(Table 5C). It is evident that the best of these three MR models is that with 15 descriptors ($S'_{cv} = 0.371$). At the same time, that model is much better than each of the NN models.

Anti-Fight Models. Multiregression models of the 53 benzodiazepines for the anti-fighting effect obtained on 104-descriptor data set are shown in Table 6A. Statistical parameters of the other models (traditional MR and NN) are shown in Table 6B. From the comparison with the NN models it is evident that the MR models with only five ($S'_{cv} = 0.345$) or six ($S'_{cv} = 0.344$) descriptors are much better than the NN models (the best model has $S'_{cv} = 0.38$). By applying the computer program *CROMRiisel.f* for the stepwise selection of descriptors in the *i by i* manner (in the same way as in the anti-pent case) starting from the best possible MR model with six descriptors, we obtained the MR models with 11, 15, and 19 descriptors. These models have even better fitting (training) values of R and S' than the NN models, but these models are not much better than the MR models with five or six descriptors according to R_{cv} and S'_{cv}.

Clined-scr Models. Only one, the best possible, multiregression model with six descriptors for the clined screen

test effect (53 compounds) is shown in Table 7A.

The six-descriptor MR model was obtained on 104-descriptor data set by the computer program *CROMRsel.f*, and possesses better fitting ($S' = 0.302$) and cross-validated ($S'_{cv} = 0.338$) statistical parameters than the best NN model ($S' = 0.31$, $S'_{cv} = 0.41$).

Number of Adjustable Parameters in the NN and MR Models. Of importance for each model is the parameter ρ—the ratio between the number of training examples and the number of adjustable parameters. It is recommended for MR models to have at least five training examples for one adjustable parameter.[41]

On the other hand, it is suggested for the NN models that ρ should be >1. The models with $\rho < 1$ are overfitted, and those with $\rho \gg$ (in practice $\rho > 2.2$) are unable to extract all relevant features and they give poor prediction.[41,42] But, despite this suggestion (which has given in 1991), to this day most authors have used NNs with $\rho \ll 1$. In our opinion, there are two possible reasons for that: (1) the authors use different variants of NNs and they believe that these rules are related only to back-propagation NN; (2) the authors believe that these rules are data set dependent and cannot be generalized on other examples, but they do not carry out such an analysis (indeed, Andrea and Kalayeh have restricted their result on "the examples of their paper"[41]). In the case of data sets used in this paper, all the considered NNs have $\rho < 1$, i.e., these NNs are overfitted according to ref 41. Overfitted NNs can be detected by the cross-validation procedure because such NNs have poor cross-validated (tested) and high fitted (trained) statistical parameters. Using cross-validated parameters as a criterion it is easily seen that the NN models from refs 34 and 35 are overfitted, but it is not the case with the NN model in ref 36. This means that the authors have avoided overfitting/overtraining problem by the averaging the predicted values calculated for analyzed cases over all network from an NN ensemble (which is composed of 500 NNs). One may expect with good reasons that MR models will easily outperform such robust NN models from refs 34 and 35 in prediction (or cross-validation) but not in fitting the data. In the studied cases we obtain such MR models that outperformed (according to calculated statistical parameters) all the NN models from refs 35 and 36 both in prediction and fitting procedures.

CONCLUSIONS

In this paper, the capabilities of multiregression models in predicting activities of four sets of carboquinones and three sets of benzodiazepines from physicochemical parameters (descriptors) are presented. The data sets of descriptors are enlarged by introduction of nonlinear cross-product terms. To find applicable MR models, we need to select a small number of significant descriptors from an enlarged data set. The use of nonlinearities between descriptors by introduction of cross-product terms is a problem also discussed by other authors.[41] However, because of the problems relating to the selection of descriptors from a large set, to the best of our knowledge, such a procedure is not used in practice.

Using procedures described in this paper, two classes of MR models can be obtained: the best possible (by *CROMRsel.f*) and the stepwise (*i by i*) models selected by the *CROMRiisel.f* program. We solved the problem of selection efficiently by using the orthogonalization procedure. The correlation coefficient of a MR model can be calculated very quickly in the orthogonal descriptor base, which enables examination of a large number of MR models. After that, several best models (or only the best one) according to high values of the correlation coefficient are selected.

By the use of programs *CROMRsel.f* or *CROMRiisel.f*, we obtained nonlinear MR models with both the fitting and predictive abilities that surpass all the published models. Effectiveness of the presented MR approach in selecting the best models as well as in introducing nonlinear relationships in MR models is demonstrated on seven data sets with about 100 descriptors.

The accuracy of each model is expressed by the training (R, S') and testing (cross-validated, R_{cv}, S'_{cv}) parameters. It is clearly seen from the comparison of the MR and NN models that the NN training parameters accessed high R and low S' values. On the other hand, values of the testing parameters for the same NN models are much lower R_{cv} and higher S'_{cv}. The large difference between R and R_{cv} or S' and S'_{cv} indicate poor predictive capabilities of the model. Namely, such a model is overfitted. This is especially so in the case of the counter-propagation[35] and back-propagation NNs.[34] However, the differences between fitting (training) and testing parameters are much lower for all the MR models (Tables 1−7), which favors the presented multiregresion approaches over neural networks and other models described in refs 34−37. Recently, Kovalishyn *et al.*[43] applied a novel robust ensemble of NNs architecture trained by the cascade-correlation learning algorithm on the carboquinone MED chronic injection data set (**MED-ci**, 37 compounds) and obtained the NN models that are inferior to the MR models obtained in the present paper. On the other hand, procedures for obtaining the MR models described in this paper that were applied on well studied Selwood data set[44,45] produced MR models that outperform even more rigorously defined NN architectures.[16] Moreover, several other NN models[16,46,47] were compared to MR models and in all studied cases MR outperforms NNs.[48]

Availability of Programs. Computer programs for the selection of the best possible descriptors (*CROMRsel.f*) and for the *i by i* stepwise selection of descriptors in multiregression models (*CROMRiisel.f*), written in FORTRAN 77, are available on request (lucic@faust.irb.hr).

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology of the Republic of Croatia through Grant 00980606. 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. We thank the reviewers for helpful comments.

REFERENCES AND NOTES

(1) Sexton, W. A. *Chemical Constitution and Biological Activity*; D. van Nostrand: New York, 1950.
(2) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.
(3) *Quantitative Structure−Activity Relationships of Drugs*; Topliss, J. G., Ed.; Academic Press: New York, 1983.
(4) *QSAR and Strategies in the Design of Bioactive Compounds*; Seydel, J. K., Ed.; VCH: Weinheim, 1985.

(5) Randić, M. Generalized Molecular Descriptors. *J. Math. Chem.* **1995**, *7*, 155−168.
(6) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.
(7) Trinajstić, N. *Chemical Graph Theory*, 2nd revised ed.; CRC Press: Boca Raton, FL, 1992; Chapter 10.
(8) 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.
(9) Randić, M. Molecular Profiles. Novel Geometry-Dependent Molecular Descriptors. *New J. Chem.* **1995**, *19*, 781−791.
(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.; 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.
(12) Hansch, C.; Fujita, T. $\rho-\sigma-\pi$ Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.
(13) Kubiny, H. *QSAR: Hansch Analysis and Related Approaches*; VCH: Weinheim, 1993; Methods and Principles in Medicinal Chemistry, Vol. 1.
(14) van de Waterbeemd, H. *Chemometric Methods in Molecular Design*; VCH: Weinheim, 1995; Methods and Principles in Medicinal Chemistry, Vol. 2.
(15) van de Waterbeemd, H. *Advanced Computer-Assisted Techniques in Drug Discovery*; VCH: Weinheim, 1995; Methods and Principles in Medicinal Chemistry, Vol. 3.
(16) 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.
(17) Camilleri, P.; Livingstone, D. J.; Murphy, J. A.; Manallack, D. T. Chiral Chromatography and Multivariate Quantitative Structure−Property Relationships of Benzimidazole Sulphoxides. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 61−69.
(18) Cramer, R. D. I.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(19) Kubinyi, H. Evolutionary Variable Selection in Regression and PLS Analyses. *J. Chemometrics* **1996**, *10*, 119−133.
(20) Rogers, D.; Hopfinger A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(21) Burns, J. A.; Whitesides, G. M. Feed-Forward Neural Networks in Chemistry: Mathematical Systems for Classification and Pattern Recognition. *Chem. Rev.* **1993**, *93*, 2583.
(22) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: An Introduction*; VCH Publishers: Weinheim, FRG, 1993.
(23) Brown, S. D.; Sum, S. T.; Despagne, F. Chemometrics. *Anal. Chem.* **1996**, *68*, 21R-61R.
(24) 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.
(25) Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, A.; Mihalić, Z. A Structure−Property Study of the Solubility of Aliphatic Alcohols in Water. *Croat. Chem. Acta* **1995**, *68*, 417−434.
(26) 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.
(27) Amić, D.; Davidović-Amić, D., Jurić, A.; Lučić, B.; Trinajstić, N. Structure−Activity Correlation of Flavone Derivatives for Inhibition of cAMP Phosphodiesterase. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1034−1038.
(28) Lučić, B.; Trinajstić, N. New Developments in QSPR/QSAR Modeling Based on Topological Indices. *SAR QSAR Environ. Res.* **1997**, *7*, 45−62.
(29) Randić, M.; Trinajstić, N. Isomeric Variations in Alkanes: Boiling Points of Nonanes. *New J. Chem.* **1994**, *19*, 179−188.
(30) Pogliani, L. Molecular Connectivity Model for Determination of Physicochemical Properties of α-Amino Acids. *J. Phys. Chem.* **1993**, *97*, 6731−6736.
(31) Pogliani, L. Molecular Connectivity Descriptors of the Physicochemical Properties of the α-Amino Acids. *J. Phys. Chem.* **1994**, *98*, 1494−1499.
(32) Hansch, C.; Hoekman, D.; Gao, H. Comparative QSAR: Toward a Deeper Understanding of Chemicobiological Interactions. *Chem. Rev.* **1996**, *96*, 1045−1075.
(33) Yoshimoto, M.; Miyazawa, H.; Nakano, H.; Shinkai, K.; Arakawa, M. Quantitative Structure−Activity Relationships in 2,5-bis(i-aziridinyl)-p-benzoquinone Derivatives Against Leukemia L-1210. *J. Med. Chem.* **1979**, *22*, 491−496.
(34) Aoyama, T.; Suzuki, Y.; Ichikawa, H. Neural Networks Applied to Quantitative Structure−Activity Relationship Analysis. *J. Med. Chem.* **1990**, *33*, 2583−2590.
(35) 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.
(36) 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.
(37) Liu, Q.; Hirono, S.; Moriguchi, I. Comparison of Functional-Link Net and Generalised Delta Rule Net in Quantitative Structure−Activity Relationship Studies. *Chem. Pharm. Bull.* **1992**, *40*, 2962−2969.
(38) Kubota, T.; Yamakawa, M.; Terada, H.; Yoshimoto, M. In *Structure−Activity Relationships-Quantitative Approaches*; the QSAR Research Group, Ed.; Kagaru no Ryoichi Suppl. Ed., No. 122, 1979.
(39) Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517−525.
(40) Szabo, A.; Ostlund, N. *Modern Quantum Chemistry*; McGraw-Hill: New York, 1989; pp 15−21.
(41) 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.
(42) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.
(43) Kovalishyn, V. V.; Tetko, I. V.; Luik, A. I.; Kholodovych, V. V.; Villa, A. E. P.; Livingstone, D. J. Neural Network Studies: Variable Selection in the Cascade-Correlation Learning Architecture. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 651−659.
(44) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure−Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, 33, 136−142.
(45) These results were presented as a poster contribution (Lučić, B.; Trinajstić, N. Multivariate Regression Excels Neural Networks, Genetic Algorithm and Partial Least-Squares in QSAR Modeling) on *12th European Symposium on Quantitative Structure−Activity Relationships: Molecular Modeling and Prediction of Bioactivity*; Copenhagen, Denmark, 23−28 August, 1998.
(46) Maddalena, D. J.; Johnston, G. A. R. Prediction of Receptor Properties and Binding Affinity of Ligands to Benzodiazepine/$GABA_A$ Receptor Using Artificial Neural Networks. *J. Med. Chem.* **1995**, *38*, 715−724.
(47) So, S.-S.; Karplus, M. Genetic Neural Networks for Quantitative Structure−Activity Relationships: Improvements and Application of Benzodiazepine Affinity for Benzodiazepine/$GABA_A$ Receptors. *J. Med. Chem.* **1996**, *39*, 5246−5256.
(48) 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, Croatia, Zagreb, 1997.

CI980090F