Toward Generating Simpler QSAR Models: Nonlinear Multivariate Regression versus Several Neural Network Ensembles and Some Related Methods

Bono Lučić,*,† Damir Nadramija,‡ Ivan Bašic,‡ and Nenad Trinajstić†

The Rugjer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia, and PLIVA, Pharmaceutical
Industry, Research Information Center, Prilaz Baruna Filipovića 25, HR-10000 Zagreb, Croatia

Received November 12, 2002

In this study we want to test whether a simple modeling procedure used in the field of QSAR/QSPR can produce simple models that will be, at the same time, as accurate as robust Neural Network Ensemble (NNE) ones. We present results of application of two procedures for generating/selecting simple linear and nonlinear multiregression (MR) models: (1) method for selecting the best possible MR models (named as CROMRsel) and (2) Genetic Function Approximation (GFA) method from the Cerius2 program package. The obtained MR models are strictly compared with several NNE models. For the comparison we selected four QSAR data sets previously studied by NNE (Tetko et al. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 794−803. Kovalishyn et al. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 651−659.): (1) 51 benzodiazepine derivatives, (2) 37 carboquinone derivatives, (3) 74 pyrimidines, and (4) 31 antimycin analogues. These data sets were parametrized with 7, 6, 27, and 53 descriptors, respectively. Modeled properties were *anti*-pentylenetetrazole activity, antileukemic activity, inhibition constants to dihydrofolate reductase from MB1428 *E. coli*, and antifilarial activity, respectively. Nonlinearities were introduced into the MR models through 2-fold and/or 3-fold cross-products of initial (linear) descriptors. Then, using the CROMRsel and GFA programs (*J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132) the sets of I ($I \leq 8$, in this paper) the best descriptors (according to the fit and leave-one-out correlation coefficients) were selected for multiregression models. Two classes of models were obtained: (1) linear or nonlinear MR models which were generated starting from the complete set of descriptors, and (2) nonlinear MR models which were generated starting from the same set of descriptors that was used in the NNE modeling. In addition, the descriptor selection method from CROMRsel was compared with the GFA method included in the QSAR module of the Cerius2 program. For each data set it has been found that the MR models have better cross-validated statistical parameters than the corresponding NNE models and that CROMRsel selects somewhat better MR models than the GFA method. MR models are also much simpler than NNEs, which is the important surprising fact, and, additionally, express calculated dependencies in a functional form. Moreover, MR models were shown to be better than all other models obtained by different methods on the same data sets ("old" multivariate regressions, functional-link-net models, back-propagation neural networks, genetic algorithm, and partial least squares models). This study also indicated that the robust NNE models cannot generate good models when applied on small data sets, suggesting that it is perhaps better to apply robust methods (like NNE ones) on larger data sets.

INTRODUCTION

Recently, we introduced a novel approach (named CROMRsel) for selecting the most important descriptors in nonlinear multiregression (MR) models.[1−4] It has also been shown that the nonlinear MR models obtained are more accurate than the models generated by the use of several robustly[1] or concisely[3] designed neural network (NN) architectures as well as the MR models obtained by the standard stepwise selection procedure used in QSAR (quantitative structure−activity relationship) modeling, like the CODESSA program.[2] All the NN models from previous comparative studies were single-NN models.

In this paper we compare linear and nonlinear MR models obtained by using the CROMRsel approach[1−3] with the models obtained by using the novel most powerful NN-based approach called the neural network ensemble (NNE).[5−7] To avoid the overfitting/overtraining problem in NN modeling Tetko et al.[5−7] have used the NNE for performing selection of the most important subset of descriptors and for the generation of the QSAR models. In their studies NNEs were composed of M networks (e.g., Tetko et al. used $M = 500$ in ref 6 and $M = 100$ in ref 7). The calculated/predicted value for each analyzed case is averaged over all M neural networks. Tetko et al. illustrated capabilities of the NNEs on four data sets, which are chosen as the standard sets for testing new methods and for performing comparative studies in the field of QSAR. Assuming the validity of Ockham's Razor in chemical modeling,[8] we continue our efforts to find simpler structure−property-activity models. Additionally, the descriptor selection method based on Genetic Function Approximation (GFA) from the QSAR module of the Cerius2 program package[9,10] was applied on the same set of data,

* Corresponding author phone: +385-1-4680-095; fax: +385-1-4680-245; e-mail: lucic@irb.hr.
† The Rugjer Bošković Institute.
‡ PLIVA.

10.1021/ci025636j CCC: $25.00
Published on Web 05/17/2003

and obtained MR models were compared with MR models obtained by CROMRsel and with NNE models. Our final result obtained on four data sets and presented in this study shows that the simplest QSAR models (obtained by CROMRsel and GFA), linear and nonlinear multiregression models, are better than the NNE models, according to the same statistical parameters.

DATA SETS AND COMPUTATIONAL METHODS

Data Sets. Four data sets were used in this study. The first data set (DS-1) contains 51 benzodiazepine derivatives with *anti*-pentylenetetrazole activity. For each molecule seven physicochemical descriptors were calculated. The second data set (DS-2) of 37 2,5-bis(1-aziridinyl)-*p*-carboquinones with antileukemic activity was parametrized with six physicochemical descriptors. The first two data sets were also previously investigated by multivariate linear regression, functional-link-net (FUNCLINK),[11,12] NNs,[12,13] and, recently, by ensemble NNs.[6,7] In refs 12 and 13 back-propagation NNs were used. Descriptors used in the present study were taken from refs 7, 12, and 13. The third data set (DS-3) consists of 74 2,4-diamino-5-(substituted benzyl)pyrimidines with inhibition constant (log K_i) of dihydrofolate reductase from MB1428 *E. coli.* Pyrimidines used were represented by 27 variables describing their physicochemical properties. This data set was also analyzed by NNs and Inductive Logic Programing[14] and by the ensemble of NNs.[6,7] The last studied data set (DS-4) is a well-known Selwood data set of 31 antimycin analogues with antfilarial activity ($-$log IC_{50}) which were parametrized with 53 physicochemical descriptors.[15] This set has often been studied by different approaches (see, e.g., ref 16 and references therein) including genetic algorithms,[10] partial least squares,[17] genetic neural networks (GNN),[18] and recently by NN ensembles.[6,7] All data sets are available on request from authors or from http://www.irb.hr/korisnici/lucic/data_sets/MRvsNNE_data_sets.

Nonlinearities were introduced into MR models through the (2- and 3-fold) cross-products of initial descriptors. Finally, to have the same initial descriptors as in the case of the NN ensemble modeling,[6,7] all the descriptors from all data sets were linearly scaled in the range between 0.1 and 0.9. This is also important because the final result in nonlinear modeling depends on the scaling procedure.[3]

Comparison Levels between MR and NNE Models. Comparison between MR models developed in this work and NNE models[6,7] was done on two main levels, depending on the initial set of descriptors used for the MR models generation: (1) comparison with linear and nonlinear MR models which were generated starting from all initial descriptors (including their cross-products) and (2) comparison with linear and nonlinear MR models which were generated starting from the subsets of descriptors selected by the NNE pruning method.[7]

In this paper, MR models of type (1) and (2) were grouped in two parts, (A) and (B), in all tables, respectively. All MR models selected by the GFA method belong to the class (1), i.e., they are selected from all initial descriptors and their 2-fold cross-products. Also, in refs 6 and 7 two sets of NNE models were obtained. The first set was based on the fixed-size back-propagation NN ensemble (BNN),[6,7] and the second set on cascade-correlation NN ensemble models (CCN).[7]

Generation of Multiregression Models by CROMRsel. MR models were generated by the use of CROMRsel procedure described previously.[1-4] All the MR models in this paper were obtained by selecting "the best possible MR models" according to the highest fitted and cross-validated correlation coefficients (i.e. we did not need to use stepwise CROMRsel selection procedures).[1]

Our selection procedure starts with the orthogonalization of descriptors, and after that, the squares of the correlation coefficients of multidescriptor models can be simply calculated (as the sums of squares of the correlation coefficients between each orthogonalized descriptor and the modeled property).[1,2] Then, leave-one-out cross-validated statistical parameters were calculated for the best model. In the case that the model selected in such a way had cross-validated parameters close to their fitted values, the selection procedure was stopped, and such a model was "proclaimed" to be the best one.

However, if the cross-validated (CV) parameters R_{cv} or S_{cv} are not very close to the corresponding fitted values (such models can be easily detected based on their fitted model solely, because they have one or more descriptors having a relatively low value of the regression coefficient comparing with its error), the model selection procedure was repeated in such a way that several best (top) models were selected. The number of these models varies in each studied case and depends on the total number of descriptors in the data set, on the level of intercorrelation of initial descriptors, and on the number of descriptors that should be selected in the models (usually, several tens or hundreds of models were checked). Then, for all of these models, the cross-validated statistical parameters were calculated. Among them, the model having the best cross-validated statistical parameters was selected as the best one.

A general scheme for performing selection of descriptors and for obtaining nonlinear MR models is as follows: (a) definition of initial size of data set $-$ definition of the number of descriptors (N); (b) generation of 2-fold and/or 3-fold cross-products of initial descriptors (for N initial descriptors, there are $N\cdot(N+1)/2!$ and $N\cdot(N+1)\cdot(N+2)/3!$ 2-fold or 3-fold cross-products, respectively), and addition of the cross-product descriptors to the initial set of N descriptors; (c) selection of the best set of I descriptors in the MR models according to the highest fitted/cross-validated correlation coefficients; and (d) computation of model parameters (regression coefficients and their errors) and other statistical parameters (fitted and cross-validated standard errors of estimate, F-test, Q^2, predictive statistical parameters for the test set compounds for which the training/test set partition is done).

All the computations have been carried out on PC computer (Athlon 1GHz).

Generation of Multiregression Models by Genetic Algorithms.[9,10] The GA perform a search over the space of possible QSAR/QSPR models using the LOF score[10] as a parameter for estimating the fitness of each model. Such evolution of a population of randomly constructed models leads to the discovery of highly predictive QSARs/QSPRs. Genetic algorithms were derived by analogy with the spread of mutations in a population. According to this analogy "individuals" are represented as a 1D string of bits. An initial population of individuals is created, usually with random

Table 1. Best Multiregression (MR) Models Containing *I* Descriptors and the Best Ensemble NN Models for 51 Benzodiazepine Derivatives Obtained for Data Set 1 (DS-1)[a]

A. MR Models Selected by the CROMRsel Program and GFA from All (Seven) Initial Descriptors and Their 2-Fold Cross-Products[b]

MR model	I	R	R_{cv}	Q^2	selected descriptors[c]
MR1-1	1	0.670	0.633	0.398	d4·d5
MR1-2	2	0.795	0.759	0.572	d1·d5, d4·d5
MR1-3	3	0.837	0.797	0.630	d1·d5, d3·d7, d4·d5
MR1-4	4	0.859	0.821	0.671	d4, d1·d4, d1·d5, d3·d7
MR1-5	5	0.880	0.839	0.700	d4, d1·d4, d1·d5, d3·d4, (d7)^2
MR1-6	6	0.896	0.854	0.725	d1, d1·d4, d1·d5, d3·d4, d4·d6, (d6)^2
MR1-7	7	0.914	0.877	0.765	d4, d6, d7, d1·d4, d1·d5, d3·d4, (d6)^2
GFA-MR1-1	4	0.859	0.819	0.671	d4, d1·d4, d1·d5, d3·d7
GFA-MR1-2	5	0.874	0.828	0.686	d4, d1·d4, d1·d5, d3·d4, d3·d7
GFA-MR1-3	6	0.886	0.828	0.686	d1·d2, d1·d5, d1·d6, d3·d7, (d4)^2, d6·d7
GFA-MR1-4	7	0.914	0.875	0.765	d4, d6, d7, d1·d4, d1·d5, d3·d4, (d6)^2

B. The Best MR Models Selected by CROMRsel Starting from Descriptors d1, d3, d4, d5, d7 (Preselected by NNE[6,7]) and Their 2-Fold Cross-Products[d]

MR model	I	R	R_{cv}	Q^2	selected descriptors
MR1-8	4	0.859	0.821	0.671	d4, d1·d4, d1·d5, d3·d7
MR1-9	5	0.880	0.839	0.700	d4, d1·d4, d1·d5, d3·d4, (d7)^2
MR1-10	7	0.891	0.850	0.720	d4, d1·d4, d1·d5, (d4)^2, d3·d4, d4·d7, d4·d5

C. The Best Ensemble NN Models[6,7]

NNE model[e]	R	R_{cv}	Q^2	descriptors used as inputs
BNN*B*-11	0.99	0.80	0.64	d1-d7
BNN*B*1-2	0.99	0.81	0.66	d1-d7
BNN1−3	0.98	0.82	0.67	d1, d3, d4, d5, d7
CCN*B*1-4		0.80	0.64	d1-d7
CCN*B*1-5		0.81	0.65	d1-d7
CCN1-6		0.82	0.66	d1, d3, d4, d5, d7

[a] Term "the best multiregression models" means the best according to the leave-one-out cross-validated correlation coefficient R_{cv}; I = the number of descriptors in multivariate regression (MR) models; R and R_{cv} = fitted and leave-one-out cross-validated correlation coefficients, respectively; Q^2 = cross-validated Q^2 value calculated by the leave-one-out procedure as is described in ref 6 (eq 14). The best MR models and the best BNN and CCN ensemble models in parts A and B of the table are underlined. [b] See Methods and ref 1 for the explanation of the CROMRsel algorithm; GFA is the Genetic Function Approximation program from Cerius2 − spline terms of initial descriptors were not used in this paper. [c] Descriptors involved in the MR model are denoted as follows (descriptors are taken from refs. 7): *d*1 = *MR*-3; *d*2 = *PI*-3; *d*3 = *MR*-7; *d*4 = σ-3; *d*5 = *F*-4; *d*6 = *R*-4; *d*7 = *I*-7. [d] These descriptors were selected by ensemble NN variable selection methods in refs 6 and 7. [e] CCN = cascade-correlation ensemble NNs (100 networks were used in the ensemble) from ref 7; BNN = cascade-correlation and fixed-size back-propagation ensemble NNs (500 networks were used in the ensemble) from ref 6.

initial bits. In the GA method, models containing a randomly chosen proper subset of the independent variables are collected, and then the collected models are "evolved". A *generation* is the set of models resulting from performing the multiple linear regression on each model. A selection of the best ones becomes the next generation (set of models). Crossover operations are performed on these, which take some variables from each of the two models to produce an offspring. In addition, the best model from the previous generation is retained. Besides linear terms, there can also be spline, quadratic, and quadratic spline terms in the QSAR module of the Cerius2 program (this module is named as GFA in Cerius2, where GFA means Genetic Function Approximation). These are added or deleted by mutation operations. However, we did not use spline, quadratic, and quadratic spline terms in this study because inclusion of such terms resulted in nonstable models (i.e. models having low values of cross-validated statistical parameters) for small data sets. A disadvantage of GFA is that it is not possible to introduce cross-products of descriptors as nonlinear terms in the GFA method included in Cerius2. Because of that, such terms are generated by the CROMRsel program and exported and used in GFA method in Cerius2. Only default values of input parameters were used in GFA (the default values for the number of generations and smoothing parameter were 5000 and 1.0, respectively). All the computations related to the models obtained by the GFA method have been carried out on SGI ORIGIN 3400.

Statistical Parameters. The quality of the models was expressed in the same way as in refs 6 and 7, i.e., by the correlation coefficient R, the cross-validated correlation coefficient R_{cv}, and the cross-validated Q^2 value. For measuring the predictive quality of the models the predictive correlation coefficient was used (the correlation coefficient between experimental and predicted activity values for compounds in the test set).

RESULTS AND DISCUSSION

Quality of the best MR models and of the best BNN[6] and CCN[7] ensemble models were evaluated by the fitted and cross-validated correlation coefficients and by the Q^2 parameter. These values are given in the corresponding Tables 1−4, together with descriptors selected in the models.

MR models selected by the CROMRsel approach are designated as MR*i*-*j*, where *i* represents the data set (i = 1−4 for data sets DS-1, ..., DS-4, respectively), and *j* denotes the (ordinal) number of model presented in corresponding Tables 1−4 for data set *i* (e.g. the name of the model MR*2*-*3* is the third (j = 3) MR model given in Table 2 developed

Table 2. Best Multiregression Models Containing *I* Descriptors and the Best Ensemble NN Models for 37 Carboquinone Derivatives Obtained for Data Set 2 (DS-2)[a]

A. MR Models Selected by the CROMRsel Program and GFA from All (Six) Initial Descriptors and Their 2-Fold Cross-Products[b]

MR model	I	R	R_{cv}	Q^2	selected descriptors[c]
MR2-1	4	0.934	0.919	0.843	d4, d5, d2•d4, d2•d6
MR2-2	6	0.955	0.937	0.877	$(d1)^2$, d1•d4, d2•d6, d3•d4, $(d4)^2$, $(d5)^2$
MR2-3	8	0.972	0.946	0.895	d2, d5, $(d1)^2$, d1•d4, d2•d4, d3•d5, $(d4)^2$, d5•d6
GFA-MR2-1	4	0.934	0.918	0.843	d4, d5, d2•d4, d2•d6
GFA-MR2-2	5	0.944	0.920	0.847	d2, d4, d5, d2•d4, $(d6)^2$
GFA-MR2-3	6	0.955	0.869	0.775	d1, d2, d4, d1•d5, d2•d4, d2•d5, d2•d6
GFA-MR2-4	7	0.958	0.925	0.856	d1, d1•d2, d2•d6, $(d3)^2$, d3•d6, d4•d6, $(d5)^2$

B. The Best MR Models Selected by CROMRsel Starting from Descriptors d2, d4, d5, d6 Preselected by NNE[6,7] and Their 2-Fold and 3-Fold Cross-Products[d]

MR model	I	R	R_{cv}	Q^2	selected descriptors[c]
MR2-4	4	0.940	0.928	0.860	d4•d6, $(d5)^2$, $(d2)^2$•d4, $(d2)^2$•d6
MR2-5	5	0.961	0.943	0.888	$(d2)^2$, d2•d4, $(d5)^2$, $(d2)^2$•d4, $(d6)^3$
MR2-6	6	0.966	0.956	0.914	d5, $(d2)^2$, d2•d4, d2•$(d4)^2$, $(d4)^3$, $(d6)^3$

C. The Best Ensemble NN Models[6,7]

NNE model[e]	R	R_{cv}	Q^2	descriptors used as inputs
BNN*B*2-1	0.99	0.89	0.79	d1-d6
BNN*B*2-2	0.98	0.92	0.84	d1-d6
BNN2-3	0.98	0.93	0.85	d2, d4, d5, d6
CCN*B*2-4		0.87	0.76	d1-d6
CCN*B*2-5		0.87	0.76	d1-d6
CCN2-6		0.88	0.78	d2, d4, d5, d6

[a,b] See footnotes a and b in Table 1. [c] Descriptors involved in the MR model are denoted as follows (descriptors are taken from refs 7 and 13): $d1 = MR_{1,2}$; $d2 = PI_{1,2}$; $d3 = PI_2$; $d4 = MR_1$; $d5 = F$; $d6 = R$. [d] These descriptors were selected by ensemble NN variable selection methods in refs 6 and 7. [e] CCN = cascade-correlation ensemble NNs (100 networks were used in the ensemble) from ref 7; BNN = cascade-correlation and fixed-size back-propagation ensemble NNs (500 networks were used in the ensemble) from ref 6.

for DS-2 ($i = 2$)). Models selected in this study by the GFA approach from the QSAR module of the Cerius2 program are designated as GFA-MR*i-j* (GFA means genetic function approximation, and the meaning of MR*i-j* is described in preceding sentence). On the other hand, NNE models are designated, in a general form, as CCN*Mi-j* or BNN*Mi-j*, where M (M = A−E in ref 6, and M = A, B, or D in ref 7) represents the pruning method used, and i and j have the same meaning as in the case of denoting the MR models. Thus, NNE models designated as BNN*D4-3* represent the third ($j = 3$) BNN based ensemble model given in Table 4, obtained by the pruning method designated with D in ref 7, which is developed for DS-4 ($i = 4$). If statistics of some other models obtained with other than MR or NNE methods are available, they will be included in the corresponding tables or given in the text.

It is important to note that in tables only the best MR models containing I ($I \leq 8$) descriptors are given. There were a lot of top MR models (with almost the same statistical parameters as the best models), which also were better than the NNE models, but they are not reported and discussed. Generally, there are two classes of the best MR models developed, and they are grouped in Tables 1−4 in two parts (A or B), depending on the initial set of descriptors used for model generation, as described in the subsection on the comparison levels between MR and NNE models.

Only for DS-4 the prediction procedure was performed for two training/test set partitions; models and their statistics are given in Table 5. In addition, regression coefficients and their errors (complete model parameters) for the best MR models from Tables 1−4 are given in Table 6.

Results for DS-1. Nonlinear MR models for 51 benzodiazepine derivatives are given in Table 1. The first set of MR models was obtained starting from all (seven) initial descriptors and from their 2-fold cross-products. They are grouped within part A of Table 1. Statistics of seven MR models (MR1-1, ..., MR1-7), that were developed starting from seven initial descriptors and 28 2-fold cross-products, are given in Table 1A. One can see that all the MR models having more than four descriptors are better than the best NNE models.

It was found in refs 6 and 7 that descriptors d2 and d6 were not relevant in NNE models and were excluded. Because of that, MR models were developed starting from d1, d3, d4, d5, and d7 as initial descriptors and their 2-fold cross-products (the MR models in Table 1B), and all are better than NNE models. Statistical parameters of the best MR models MR*1-7* ($Q^2 = 0.765$) and MR*1-10* ($Q^2 = 0.720$) are better than corresponding parameters of the best NNE models, both BNN ($Q^2 = 0.67$) and CCN ($Q^2 = 0.66$) ensembles.

It is important to note that the best nonlinear MR models having 1−5 descriptors do not include descriptors d2 and d6, at all. However, the MR model obtained by CROMRsel containing six descriptors (MR1-6) does not contain descriptors d2 and d7, but the next model MR1-7 contains all descriptors except d2. Moreover, descriptor d2 is not included in the best GFA models containing 4, 5, and 7 descriptors (Table 1A). Therefore, one can see that, for MR based models, only descriptor d2 is irrelevant descriptor. In addition, it is evident that all the reported models not only are nonlinear, including mostly 2-fold cross-products, but also some involve linear terms (initial descriptors). Details of the MR models MR1-6 (Table 1A) and MR1-9 (Table 1B) are given in Table 6. It is also shown that MR models are better than the FUNCLINK ($R = 0.88$, $Q^2 = 0.71$)[6,12]

Table 3. Best Multiregression Models Containing *I* Descriptors and the Best Ensemble NN Models for 74 Pyrimidines Obtained for Data Set 3 (DS-3)[a]

A. The Best MR Models Selected by the CROMRsel Program and GFA[b]

MR model	*I*	*R*	R_{cv}	Q^2	selected descriptors[c]
					From All (27) Descriptors
MR3-1	6	0.868	0.801	0.638	d2, d3, d11, d12, d20, d23
MR3-2	7	0.878	0.811	0.654	d2, d3, d4, d11, d12, d20, d23
GFA-MR3-1	6	0.868	0.799	0.638	d2, d3, d11, d12, d20, d23
GFA-MR3-2	7	0.881	0.689	0.474	d2, d3, d17, d19, d22, d24, d26
					From All (27) Descriptors and Their 2-Fold Cross-Products
MR3-3	2	0.863	0.849	0.714	d5·d24, d6·d20
MR3-4	3	0.912	0.895	0.801	d1·d20, d1·d23, d8·d17
MR3-5	4	0.928	0.906	0.820	d1·d20, d1·d23, d4·d12, d8·d17
GFA-MR3-3	2	0.621	0.563	0.317	d6·d17, d15·d20
GFA-MR3-4	3	0.895	0.824	0.679	d2·d17, d9·d22, d9·d26
GFA-MR3-5	4	0.921	0.902	0.813	d4·d24, d8·d11, d9·d20, $(d12)^2$

B. The Best MR Models Selected by CROMRsel Starting from Descriptors Preselected by NNE[7,d]

MR model	*I*	*R*	R_{cv}	Q^2	selected descriptors
					d2, d3, d11, d12, d20, d23 and Their 2-Fold Cross-Products[d]
MR3-6	4	0.863	0.834	0.694	d20, d23, d2·d23, d3·d20
MR3-7	5	0.879	0.849	0.720	d20, d23, d2·d23, d3·d20, d12·d20
MR3-8	6	0.905	0.876	0.767	d11, d20, d23, d2·d23, d3·d20, $(d11)^2$
MR3-9	7	0.910	0.894	0.798	d20, d2·d11, $(d3)^2$, d3·d20, d3·d23, d11·d23, $(d23)^2$
					d2-d4, d8, d11, d12, d15, d20, d23 and Their 2-Fold Cross-Products[d]
MR3-10	4	0.874	0.846	0.714	d8·d11, d11·d12, d20·d23, $(d23)^2$
MR3-11	5	0.890	0.867	0.746	d4·d12, d4·d23, d8·d15, d20·d23, $(d23)^2$
					d2-d4, d6, d8, d11, d12, d17, d20, d23 and Their 2-Fold Cross-Products[d]
MR3-12	3	0.896	0.872	0.739	d6·d20, d6·d23, d8·d17
MR3-13	4	0.925	0.910	0.824	d6·d20, d6·d23, d8·d11, $(d12)^2$

C. The Best Ensemble NN Models[7]

NN model[e]	*R*	R_{cv}	Q^2	descriptors used as inputs
BNN3-1		0.63	0.40	all descriptors 1−27
CCN3-2		0.62	0.37	all descriptors 1−27
BNN*A*3-3		0.79	0.63	d2, d3, d4, d6, d8, d11, d12, d17, d20, d23
CCN*A*3-4		0.78	0.62	d2, d3, d4, d6, d8, d11, d12, d17, d20, d23
BNN*B*3-5	-	0.81	0.65	d2, d3, d11, d12, d20, d23
CCN*B*3-6		0.83	0.68	d2, d3, d11, d12, d20, d23
BNN*D*3-7	-	0.80	0.64	d2, d3, d4, d8, d11, d12, d15, d20, d23
CCN*D*3-8		0.82	0.68	d2, d3, d4, d8, d11, d12, d15, d20, d23

[a,b] See footnotes a and b in Table 1. [c] Descriptors involved in the MR model correspond to those (the number of columns) in the data file from refs 7 and 14. [d] These descriptors were selected by ensemble NN variable selection methods in ref 7. [e] CNN = cascade-correlation ensemble NNs (100 networks were used in the ensemble) from ref 7; BNN = cascade-correlation and fixed-size back-propagation ensemble NNs were trained as described in ref 6.

and GDR-NN (generalized delta rule neural network, $R = 0.865$, $R_{cv} = 0.566$)[6,12] models.

Results for DS-2. Nonlinear MR models developed on 37 carboquinone derivatives are given in Table 2. The first set of MR models was developed starting from six initial descriptors and their 2-fold cross-products. Statistical parameters of MR models containing four, six, and eight descriptors are given in Table 2A. These models are highly nonlinear with a lot of 2-fold cross-products (squares are also treated as cross-products) of initial descriptors. There is also evidence that the most irrelevant, according to the CROMRsel and GFA, is descriptor d3. According to both NNE approaches (BNN and CCN) the irrelevant descriptors are descriptors d1 and d3.

Taking this into account, nonlinear MR models starting from descriptors d2, d4, d5, and d6 with their 2- and 3-fold cross-products are selected by CROMRsel and given in Table 2B. Even the four-descriptor MR model (MR2-4, $Q^2 = 0.86$) is better than the best NNE model (BNN-3, $Q^2 = 0.85$). Moreover, the MR model containing six descriptors (MR2-6) has $Q^2 = 0.914$ ($R = 0.966$, $R_{cv} = 0.956$). This model is also better than the best FUNCLINK ($R = 0.95$, $Q^2 = 0.87$)[6,12] and GDR-NN ($R = 0.94$, $R_{cv} = 0.86$)[6,12] models. Regression coefficients and their errors for this model are also given in Table 6.

Results for DS-3. The quality of the models, developed on the third set containing 74 pyrimidines and parametrized by 27 descriptors, should be much more dependent on the selection of descriptors, due to the larger descriptor set. Two linear MR models obtained by CROMRsel (MR3-1 and MR3-2) and two linear models obtained by GFA (GFA-MR3-1 and GFA-MR3-2) are given in Table 3A. Even these linear models have statistical characteristics comparable to the corresponding best NNE models (BNN3-1 and CCN3-2 from Table 3C), which are, inherently, nonlinear models. Then, to initial descriptors their cross-products are added, and the best MR models selected by CROMRsel containing 2, 3, and 4 descriptors are selected. Each of these models (MR3-3, MR3-4, and MR3-5) is better than the best NNE and GFA based models. In addition, nonlinear MR model

Table 4. Best Multiregression Models Containing *I* Descriptors and the Best Ensemble NN Models for 31 Antimycin Analogues Obtained for Data Set 4 (DS-4)[a]

A. The Best MR Models Selected by the CROMRsel Program and GFA[b]

MR model	I	R	R_{cv}	Q^2	selected descriptors[c]
From 53 Descriptors					
MR4-1	2	0.781	0.732	0.533	d39, d50
MR4-2	3	0.849	0.806	0.647	d38, d50, d52
MR4-3	4	0.863	0.817	0.665	d12, d38, d50, d52
MR4-4	5	0.904	0.842	0.699	d4, d11, d39, d50, d52
MR4-5	6	0.924	0.871	0.754	d4, d11, d38, d48, d50, d52
GFA-MR4-1	2	0.690	0.570	0.325	d35, d50
GFA-MR4-2	3	0.781	0.709	0.502	d39, d49, d50
GFA-MR4-3	4	0.877	0.788	0.621	d4, d5, d38, d50
GFA-MR4-4	5	0.909	0.834	0.696	d4, d5, d11, d39, d50
GFA-MR4-5	6	0.924	0.867	0.751	d4, d11, d39, d48, d50, d52
GFA-MR4-6	7	0.937		−0.021	d4, d5, d24, d27, d30, d39, d50
From 53 Descriptors and Their 2-Fold Cross-Products					
MR4-6	2	0.871	0.834	0.693	d4•d51, d5•d50
MR4-7	3	0.926	0.897	0.802	d4•d50, d5•d50, d40•d50
GFA-MR4-7	2	0.869	0.830	0.689	d1•d50, d14•d35
GFA-MR4-8	3	0.907	0.865	0.749	d4•d5, d11•d51, d50•d51
GFA-MR4-9	4	0.921	0.884	0.781	d1•d50, d4•d51, d8•d14, d47•d49
B. The Best MR Models Selected by CROMRsel Starting from Descriptors Preselected by NNE[6,d]					
d24, d50, d51 and Their 2-Fold Cross-Products[d]					
MR4-8	2	0.757	0.704	0.492	d24•d51, d50•d51
MR4-9	4	0.811	0.743	0.545	d24•d50, d24•d51, d50•d51, (d51)^2
d11, d13, d14, d35, d50, d52 and Their 2-Fold Cross-Products[d]					
MR4-10	3	0.882	0.854	0.727	d50, d35•d50, d35•d52
MR4-11	4	0.914	0.881	0.771	d50, d11•d13, d35•d50, d50•d52
d4, d6, d11, d13, d14, d35, d50, d52 and Their 2-Fold Cross-Products[d]					
MR4-12	2	0.829	0.786	0.614	d4•d50, d6•d35
MR4-13	4	0.921	0.891	0.792	d50, d4•d52, d11•d13, d35•d50
d4, d6, d7, d11, d13, d14, d35, d50, d52 and Their 2-Fold Cross-Products[d]					
MR4-14	3	0.882	0.854	0.727	d50, d35•d50, d35•d52
MR4-15	4	0.921	0.891	0.792	d50, d11•d13, d35•d50, d50•d52

C. The Best Ensemble NN Models[7]

NN model[e]	R	R_{cv}	Q^2	descriptors used as inputs
BNN4-1		0.57	0.32	all descriptors 1−53
CCN4-2		0.57	0.32	all descriptors 1−53
BNN4-3		0.66	0.43	d24, d50, d51
CCN4-4		0.68	0.46	d24, d50, d51
BNN*A*4-5		0.82	0.67	d11, d13, d14, d35, d50, d52
CCN*A*4-6		0.80	0.64	d11, d13, d14, d35, d50, d52
BNN*B*4-7		0.82	0.67	d4, d6, d11, d13, d14, d35, d50, d52
CCN*B*4-8		0.82	0.67	d4, d6, d11, d13, d14, d35, d50, d52
BNN*D*4-9		0.81	0.66	d4, d6, d7, d11, d13, d14, d35, d50, d52
CCN*D*4-10		0.81	0.66	d4, d6, d7, d11, d13, d14, d35, d50, d52

D. The Best Models Obtained by Other Methods for DS-4

model[f]	R	R_{cv}	Q^2	descriptors used as inputs
GNN[18]	0.919	0.866		d27, d38, d50
GA[10]	0.920	0.849		d4, d5, d6, d11, d39, d50
PLS[17]	0.910		0.694	d4, d5, d11, d17, d36, d38, d39, d40, d50, d52

[a,b] See footnotes a and b in Table 1. [c] Descriptors involved in the MR model are ordered in the same way as in ref 7. [d] These descriptors were selected by ensemble NN variable selection methods in ref 7. [e] CNN = cascade-correlation ensemble NNs (100 networks were used in the ensemble) from ref 7; BNN = cascade-correlation and fixed-size back-propagation ensemble NNs from ref 7. BNN were trained as described in ref 6. [f] GNN = genetic neural network; GA = genetic algortihm; PLS = partial least squares.

GFA-MR3-5, obtained by GFA ($Q^2 = 0.813$), is also better than the best NNE model (CCN*B*3-6, $Q^2 = 0.68$).

To improve NNE models, the authors selected smaller subsets of descriptors which contain all (or almost all) of the necessary information. Using the NNE pruning method they selected three subsets containing 10, 6, and 9 the most significant descriptors. Starting from each subset and including cross-products of initial descriptors, the best nonlinear MR models for each subset were selected by CROMRsel. These MR models and corresponding NNE models are given in Table 3B and 3C. Again, each of the MR models given in Table 3B is better then the best NNE models.

Results for DS-4. This is the largest analyzed data set according to the total number of initial descriptors. The power

Table 5. Comparison between Multiregression Models and Ensemble NN Models for 31 Antimycin Analogues (Data Set 4) Using Training/Test Sets Protocols[a]

A. Multiregression Models Selected by the CROMRsel Program[b]

From 53 Descriptors

	training set (1−16)		test set (17−31)	selected
I	R_{cv}	Q^2	R	descriptors[c]
2	0.859	0.738	0.719	d50, d51
3	0.902	0.797	0.718	d22, d50, d51
4	0.940	0.883	0.778	d4, d5, d50, d51

	training set (17−31)		test set (1−16)	selected
I	R_{cv}	Q^2	R	descriptors[c]
2	0.815	0.656	0.623	d38, d50

From 53 Descriptors and Their 2-Fold Cross-Products

	training set (1−16)		test set (17−31)	selected
I	R_{cv}	Q^2	R	descriptors[c]
1	0.868	0.753	0.604	d50•d51
2	0.931	0.866	0.697	d4•d5, d50•d51
From Descriptors d4, d6, d13, d14, d35, d46, d50, d51 and Their 2-Fold and 3-Fold Cross-Products[d]				
2	0.909	0.821	0.691	d4•d6•d51, d4•d50•d51
2	0.927	0.855	0.693	d6•d13, d4•d50•d51
3	0.966	0.932	0.725	d4•(d50)^2, d6•d13•d14, d13•d46•d51
From Descriptors d4, d6, d50, d51 and Their 2-Fold and 3-Fold Cross-Products[d]				
2	0.909	0.821	0.691	d4•d6•d51, d4•d50•d51
4	0.910	0.827	0.838	(d4)^2, (d6)^2, (d4)^2•d51, d50•(d51)^2
From Descriptors d4, d6, d13, d14, d46, d50, d51 and Their 2-Fold and 3-Fold Cross-Products[d]				
2	0.909	0.821	0.691	d4•d6•d51, d4•d50•d51
2	0.927	0.855	0.693	d6•d13, d4•d50•d51
3	0.966	0.932	0.725	d4•(d50)^2, d6•d13•d14, d13•d46•d51

From Descriptors d29, d33, d50[d]

	training set (17−31)		test set (1−16)	selected
I	R_{cv}	Q^2	R	descriptors[c]
3	0.676	0.40	0.665	d29, d33, d50

B. The Best Ensemble NN Models from Ref 7

	training set (1−16)		test set (17−31)	descriptors
method[e]	R_{cv}	Q^2	R	used as inputs
CNN	0.87	0.72	0.65	d4, d6, d13, d14, d35, d46, d50, d51
CNN	0.87	0.74	0.75	d4, d6, d50, d51
CNN	0.88	0.72	0.68	d4, d6, d13, d14, d46, d50, d51

	training set (17−31)		test set (1−16)	descriptors
method[e]	R_{cv}	Q^2	R	used as inputs
CNN	0.73	0.52	0.56	d29, d33, d50

[a,b] See footnotes a and b in Table 1. [c,d] See footnotes c and d in Table 4. [e] CNN = cascade-correlation ensemble NNs (100 networks were used in the ensemble) from ref 7.

Table 6. Details of Several the Best Models from Tables 1−4 Selected by CROMRsel

MR model	equation
MR1-6	**anti-pent** = 5.69 (±0.20) − 0.64 (±0.18) d1−6.28 (±1.0) d1•d4 + 7.85 (±1.54) d1•d5 − 1.92 (±0.42) d3•d4 + 5.01 (±0.42) d4•d6 − 3.81 (±0.39) (d6)^2
MR1-9	**anti-pent** = 3.99 (±0.19) + 3.21 (±0.29) d4 − 6.43 (±1.0) d1•d4 + 7.32 (±1.56) d1•d5 − 1.74 (±0.45) d3•d4 − 0.63 (±0.19) (d7)^2
MR2-6	**OD-ci** = 7.22 (±0.14) − 1.86 (±0.23) d5 − 1.83 (±0.44) (d2)^2 − 10.72 (±1.54) d2•d4 + 16.52 (±1.80) d2•(d4)^2 −3.15 (±0.36) (d4)^3 −1.28 (±0.15) (d6)^3
MR3-3	**log K_i** = 0.6177 (±0.097) − 1.35 (±0.11) d5•d24 + 1.65 (±0.12) d6•d20
MR3-4	**log K_i** = 0.575 (±0.010) + 2.22 (±0.16) d1•d20 − 1.89 (±0.14) d1•d23 − 0.419 (±0.068) d8•d17
MR3-5	**log K_i** = 0.5920 (±0.010) + 2.11 (±0.15) d1•d20 − 1.69 (±0.14) d1•d23 − 0.88 (±0.23) d4•d12 + 0.441 (±0.062) d8•d17
MR3-13	**log K_i** = 0.5875 (±0.0089) + 1.48 (±0.11) d6•d20 − 1.245 (±0.090) d6•d23 + 0.549 (±0.087) d8•d11 − 0.247 (±0.040) (d12)^2
MR4-6	**-log(IC_{50})** = −1.41 (±0.22) + 1.91 (±0.38) d4•d51 + 3.83 (±0.50) d5•d50
MR4-7	**-log(IC_{50})** = −1.36 (±0.17) + 4.67 (±0.65) d4•d50 + 3.86 (±0.45) d5•d50 − 4.47 (±0.60) d40•d50
MR4-11	**-log(IC_{50})** = −1.05 (±0.32) + 7.4 (±1.1) d50 − 1.82 (±0.54) d11•d13 −8.5 (±1.1) d35•d50 + 3.66 (±0.63) d50•d52
MR4-13	**-log(IC_{50})** = −1.54 (±0.29) + 7.7 (±1.0) d50 + 2.84 (±0.46) d4•d52 − 2.04 (±0.53) d11•d13 − 7.4 (±1.0) d35•d50

of the CROMRsel approach (and GFA, as well) to the selection of descriptors for MR models is best exemplified in this case. The first set of the models was developed starting from all (53) descriptors and the second starting from 53 initial descriptors and their cross-products. These models are given in Table 4A. Comparing statistical parameters of these MR models with the best NNE models (Table 4C, Q^2 = 0.67) one can see that even the linear five-descriptor MR models (Table 4A: MR4-4, Q^2 = 0.699, and GFA-MR4-4, Q^2 = 0.696) are better. Taking into account initial descriptors and their 2-fold cross-products we obtained two-descriptor (MR4-6, Q^2 = 0.693) and three-descriptor (MR4-7, Q^2 = 0.802) models selected by CROMRsel which are better than the corresponding GFA and NNE models. In addition, nonlinear models selected by GFA having 2−4 descriptors are also better than the corresponding NNE ones.

Comparison between MR models (Table 4B) selected by CROMRsel and the corresponding NNE models (Table 4C) for the same initial subsets of descriptors (which are preselected by the NNE pruning procedure) clearly shows that all MR models are better than the NNE ones. In this case, for example, models MR4-8 and MR4-9 should be compared with BNN4-3 and CCN4-4 models. The best MR models given in Table 4 (MR4-5, MR4-7, MR4-11, MR4-13, MR4-15) are also better than the best GNN (genetic neural network) (R = 0.919, R_{cv} = 0.866),[18] partial least squares (R = 0.910, Q^2 = 0.694),[17] and the genetic algorithm based (R = 0.920, R_{cv} = 0.849) models.[10] A lot of different methods were applied on the Selwood data set more or less successfully (see e.g. ref 16 and references cited therein, and ref 19 and refs 18−30 cited therein). The latest two applications of variable selection algorithms (unsupervised forward selection[19] and fast random elimination of descriptors (FRED)[20]) gave for the best models $Q^2 \approx 0.5$ and Q^2 = 0.683, respectively. One can easily see that results obtained for DS-4 in this study (Table 4) are better.

One point is worth mentioning. If we plot in Figure 1 R vs R_{cv} for all linear six-descriptor models with $R > 0.89$ selected by CROMRsel (there are 3329 such models), one can see that there are great many (420) linear MR models beyond the line at R_{cv} = 0.82 (denoting the best NNE model). The same case was also observed for all other studied data

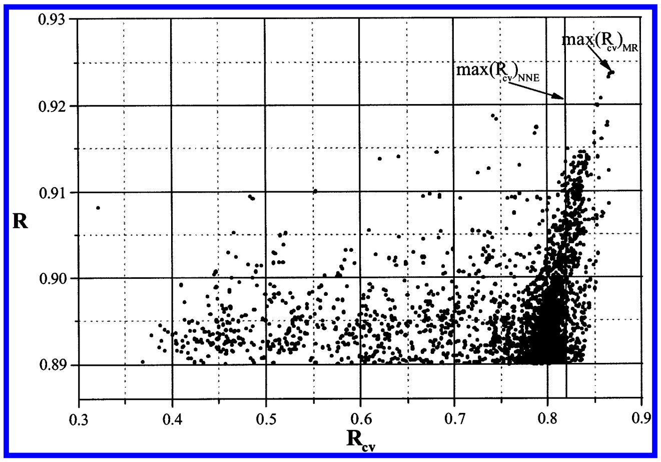

Figure 1. *R* versus R_{cv} plot of the 3329 top linear six-descriptor MR models with $R > 0.89$ selected by CROMRsel from 53 descriptors (DS-4).

sets and for other selected models, i.e., there were a lot of linear MR models having values of (both fitted and CV) statistical parameters between those for the best NNE and those for the best MR models.

The selection of descriptors based on CROMRsel (for linear models) recognized descriptors d50, d52, d38, d4, and d11 as the most important. This agrees with selections of descriptors done by other methods. However, it appears that the most important descriptor in the best linear and nonlinear MR models obtained both by CROMRsel and GFA is logP (d50, hydrophobicity parameter = logarithm of the octanol/water partition coefficient). This parameter is the best parameter for modeling the drug transport through membrane.

Prediction Test Results for DS-4. In ref 7 the authors performed training/test set partition of the DS-4 in order to see the predictive quality of the NNE models. Mostly, they used compounds 1−16 for training and compounds 17−31 for testing the model quality. Only in two cases compounds 17−31 were used for training and compounds 1−16 for testing the models. Quality of prediction was measured by the predictive correlation coefficient. We followed, step by step, the NNE procedure and developed the corresponding linear/nonlinear MR models by CROMRsel, starting from the same subset of initial descriptors (which are selected by NNE and given in ref 7). In this case (training/test set partition) GFA were not used because such a procedure cannot be automatically performed and validated in GFA from Cerius2. MR models are given in Table 5A, and the corresponding NNE results are in Table 5B. It is interesting to note that in all models descriptor d50 is involved (log P), either as a linear descriptor or as a part of 2- or 3-fold cross-products. A weak point of the results given in Table 5 is that the training/test set partitions produce a very small training data set (only 15 or 16 compounds). For such a small data set it is not easy to obtain good and stable models. This is best seen from the (sometimes) very large difference between the correlation coefficients for the training set and for the test set.

CONCLUSION

This study is important due to the growing interest in application of neural networks to quantitative structure−activity relationships. In our opinion, results reported here should be considered by all researchers engaged in the QSAR modeling suggesting that they should make efforts to obtain simpler models. We believe that NN and NNE modelers should especially consider results obtained and presented here and try to improve NN and NNE modeling and related models in the future, taking into account presented results for MR models.

Additionally, easily interpretable models should be preferred over the complex ones. This is more easily achieved if one uses the simple functional form of models relating structural parameters (descriptors) with the property/activity of molecules. As it is shown in this study, the linear or nonlinear multiregressions, as the simplest modeling procedures giving the simplest functional form of models, together with an efficient procedure for selection of the most important descriptors, can produce very concise and very good models. It came out from the analyses of these four standard data sets that the best MR models are much simpler than the best NNE. The best MR models selected by CROMRsel and GFA contain 3−9 optimized parameters (including constant term). On the other hand, the best NNE models contains several hundreds/thousands of optimized parameters (weights).[21] Surprisingly, such drastic simplifications through reduction of the number of optimized parameters were obtained by using CROMRsel and GFA descriptor selection approaches with, at the same time, improvement of the statistical performances of selected MR models.

Comparison between CROMRsel and GFA descriptor selection methods show that CROMRsel produced more stable models on these four data sets. However, in some cases we selected by GFA as good models as by CROMRsel. This is in agreement with our recent results obtained on viscosity modeling.[22] Additionally, MR models selected by GFA were, almost in all cases, better than those selected by NNE.

Moreover, it is also shown on the most often used Selwood data set that the best MR models obtained by the CROMRsel program are better than previously published models obtained by using genetic algorithms,[10] partial least squares,[17] or genetic neural networks.[18]

ACKNOWLEDGMENT

We are grateful to V. V. Kovalishyn and I. V. Tetko for providing benzodiazepine, carboquinone, and pyrimidine data sets and to A. J. Hopfinger and D. Rogers for providing Selwood data set. This work was supported by the Ministry of Science and Technology of the Republic of Croatia through Grants 0098034 (N.T. and B.L.).

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) 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.

(3) Lučić, B.; Amić, D.; Trinajstić, N. Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 403−413.

(4) Basak, S. C.; Gute, B. D.; Lučić, B.; Nikolić, S.; Trinajstić, N. A Comparative QSAR Study of Benzamidines Complement-Inhibitory Activity and Benzene Derivatives Acute Toxicities. *Comput. Chem.* **2000**, *24*, 181−191.

(5) Tetko, I. V.; Alessandro Villa, A. E. P.; Livingston, D. J. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 826−833.

(6) 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.

(7) Kovalishyn, V. V.; Tetko, I. V.; Alessandro Villa, A. E. P.; Livingston, D. J. Neural Network Studies. 3. Variable Selection in the Cascade-Correlation Learning Architecture. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 651−659.

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

(9) Cerius2; Accelrys: 9685 Scranton Road, San Diego, CA, 92191.

(10) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 854−866.

(11) Nakao, H.; Arakawa, M.; Nakamura, T.; Fukushima, M. Antileukemic Agents. II. New 2,5-bis(1-aziridinyl)-*p*-benzoquinone Derivatives. *Chem. Pharm. Bull*. **1972**, *20*, 1968−1979.

(12) 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.

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

(14) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative Structure−Activity Relationships by Neural Networks and Inductive Logic Programming. 1. The Inhibition of Dihydrofolate Reductase by Pyrimidines. *J. Comput.-Aided Mol. Des*. **1994**, *8*, 405−420.

(15) 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.

(16) Dunn, W.; Rogers, D. Genetic Partial Least Squares in QSAR. In *Genetic Algorithms in Molecular Modeling*; Devillers, J., Ed.; Academic Press: London, 1996; pp 109−130.

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

(18) 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.

(19) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1160−1168.

(20) Waller, C. L.; Bradley, M. P. Development and Validation of a Novel Variable Selection Techniques with Application to Multidimensional Quantitative Structure−Activity Relationships Studies. *J. Chem. Inf. Comput. Sci*. **2000**, *39*, 345−355.

(21) For example, in ref 6 the authors pointed on page 795 (section "ANN implementation") that single ANN (Artificial Neural Network) with 10 neurons in one hidden layer were used in the computation. In addition, bias neurons were used both on the input and on the hidden layer. It means that single ANN contained up to 77 weights for only five inputs. Because at least 100 ANNs were calculated and used in each ensemble a huge number of more than 7000 weights (i.e. optimized parameters) were used in each NNE model. It is important to point out that, for some data sets, even 500 ANNs were used in the ensemble.

(22) Lučić, B.; Bašic, I.; Nadramija, D.; Miličević, A.; Trinajstić, N.; Suzuki, T.; Petrukhin, R.; Karelson, M.; Katritzky, A. R. Correlation of liquid viscosity with molecular structure for organic compounds using different variable selection methods. *Arkivoc* **2002**, (IV), 45−59 (http://www.arkat-usa.org/ark/journal/2002/Sunko/DS-381D/DS-381D.pdf).

CI025636J