# Kohonen Artificial Neural Network and Counter Propagation Neural Network in Molecular Structure-Toxicity Studies

Marjan Vracko*

*National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia*

**Abstract:** We present self-organizing map or Kohonen network and counter propagation neural network as powerful tools in quantitative structure property/activity relationship modeling. Two areas of applications are discussed: estimation of toxic properties in environmental research and applications in drug research.

**Keywords:** SOM, Kohonen neural network, counter propagation neural network, QSAR.

## INTRODUCTION

In this article we give an overview over the using of self-organizing maps (SOM) and counter propagation artificial neural networks (CP NN) in Quantitative Structure-Activity/Toxicity Relationship studies (QSAR/QSTR). Some specifics, which make mentioned techniques attractive in QSAR/QSPR are discussed. In comparison to other neural network methods the SOM and CP NN are transparent and the results are easily interpretable. The basic result of SOM is an arrangement of objects in a two-dimensional network on the basis of similarity among objects. One of the reasons to perform this mapping is the visualization of data. Our mind can not analyze the data in multi-dimensional spaces, but it is very effective in analyzing of two-dimensional pictures. From two-dimensional pictures one easily recognizes similarity relationships among the objects. The details, i.e. the architecture of network, learning strategy, basic properties, and how the SOM and CP NN are used in QSPR/QSAR are described in the sections below.

An interesting area of application is estimation of toxic properties of chemicals for risk assessment and classification for regulatory purposes. There are more than 82000 substances in current use as chemicals in commerce in USA and EU. According to the Toxic Substances Control Act (TSCA) Inventory this number is growing by nearly 3000 substances per year [1]. Only about 15% of TSCA chemicals have data necessary for reliable risk assessment. Similar situation is on European market. Institute for Health and Consumer Protection reports that different toxicity data are available for 15% - 70% of High Production Volume Chemicals [2]. For the estimation of missing data, which are required for risk assessment of chemicals the using of QSAR and similar computational methods is recommended from producers and regulatory bodies. The theoretical approach faces several problems about the consistency of databases, representation of molecular structures, building of models, and testing and validation of models. The databases usually consist of diverse chemicals, i.e., the chemicals, which are active due to different mechanisms. The theoretical methods should be robust enough to treat this diversity within a single model. Mentioned techniques seem to be suitable for this.

Other field of interest is drug research. The QSAR employing mostly linear regression method as a modeling technique plays an important role in drug design, particularly in the latter stage of drug development when the lead compound is already known [3]. The role of the standard QSAR approach is to find in the set of congeneric compounds the optimal drug. In the last years new aspects in drug development are approaching: to find new leads acting on the same receptor or to find drugs acting on different receptors but having the same biological effect. For these purposes diverse data sets must be considered, which can be treated more effectively with non-linear methods [4, 5].

A general problem in QSAR modeling is the selection of most relevant descriptors. Descriptors are parameters calculated from a molecular structure, or they are measured with different physico-chemical methods. They represent a molecular structure in a model, or with the other words they encode the molecular structure on numerical way. Nowadays hundreds of descriptors are in use. Further informations on their systematic, methods of calculation and application are given in References [6, 7]. In last decade the methods like neural networks including SOM, genetic algorithm and fuzzy logic methods are often used to select important descriptors [8, 9].

## ARCHITECTURE AND LEARNING STRATEGY OF SOM AND CP NN

SOM or Kohonen neural network is one of the basic types of artificial neural networks. Its architecture represents a two-dimensional grid of connected neurons, which are multi-dimensional vectors. (The dimension of vectors is equal to the number of descriptors.) The learning of SOM is the projection from multi-dimensional space onto two-dimensional grid (array) of neurons. The projection or learning of network runs in two-steps, the first step is the selection of the winning neuron and the second step is the self-organization of the map [10]. In details it runs as follows. A vector, which represents an object is presented to all neurons and the algorithm selects the neuron that is most similar to it (winning neuron). In the second step the weights of the winning neuron are modified to the vector

*Address correspondence to this author at the National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia; Tel.: +386 1 4760315, Fax: +386 1 4760300, E-mail: marjan.vracko@ki.si

1573-4099/05 $50.00+.00

values and in the same time the neighboring neurons are modified to become similar to it. Details and mathematical expressions are discussed in several textbooks and articles [8, 9]. After all objects are presented to the network one learning epoch is over. This procedure repeats until the weights are stabilized. As a result one obtains objects organized in two-dimensional map with layer structure, where each layer represents one component of multi-dimensional vector (one descriptor). The mapping is topology preserving what means that similar objects in descriptor space are located close to each other (or even on the same neuron) but it is not metric preserving. There are attempts to include the metric information into SOM, for example the using of colored graphs [11]. Indeed, the lost of metric information is not a shortcoming. The projection from multi-dimensional space onto very limited grid of neurons caused overlapping and squeezing of information. With the other words, a region in two-dimensional grid carries information from different regions of original space. A map is not only a picture of original space, but also a model. In this stage of training only input variables (representation vectors) were taken into account and therefore the SOM is referred as unsupervised network.

The simplest way to include the output variables (property values) is to increase the dimension of neurons and treat the input and output variables equivocally [12]. The counter propagation neural network (CP NN) implements input and output variables differently [9, 13]. The architecture of CP NN is shown in Fig. (**1**). It has two layers the input layer, which has the same structure as in SOM and the output layer situated beneath. The difference to SOM lies in the learning strategy. The learning in the input layer is the same as in SOM, i.e., the input variables determinate the arrangement of objects. When the arrangement is set the positions of objects are projected to the output layer where the weights are modified in such a way that the weights on projected positions correspond to the output values. In addition, the weights in the neighborhood are modified. On this way the response surface is constructed. This part of training is conducted considering the output values and therefore it is usually referred as the supervised part of training of CP NN. Similarly, the prediction runs over two steps. In the first step the object is located into input layer on the neuron with the most similar weights. In the second step, the position of that neuron is projected to the output layer, which gives the predicted output value.

## APPLICATIONS IN QSAR MODELING

### Modeling

In comparison to other neural networks the SOM and CP NN have transparent structure, i. e., the results can be easily interpreted. An advantage is that we can follow the predictions. When an object is situated into the trained network its neighbors determine the prediction. In addition,

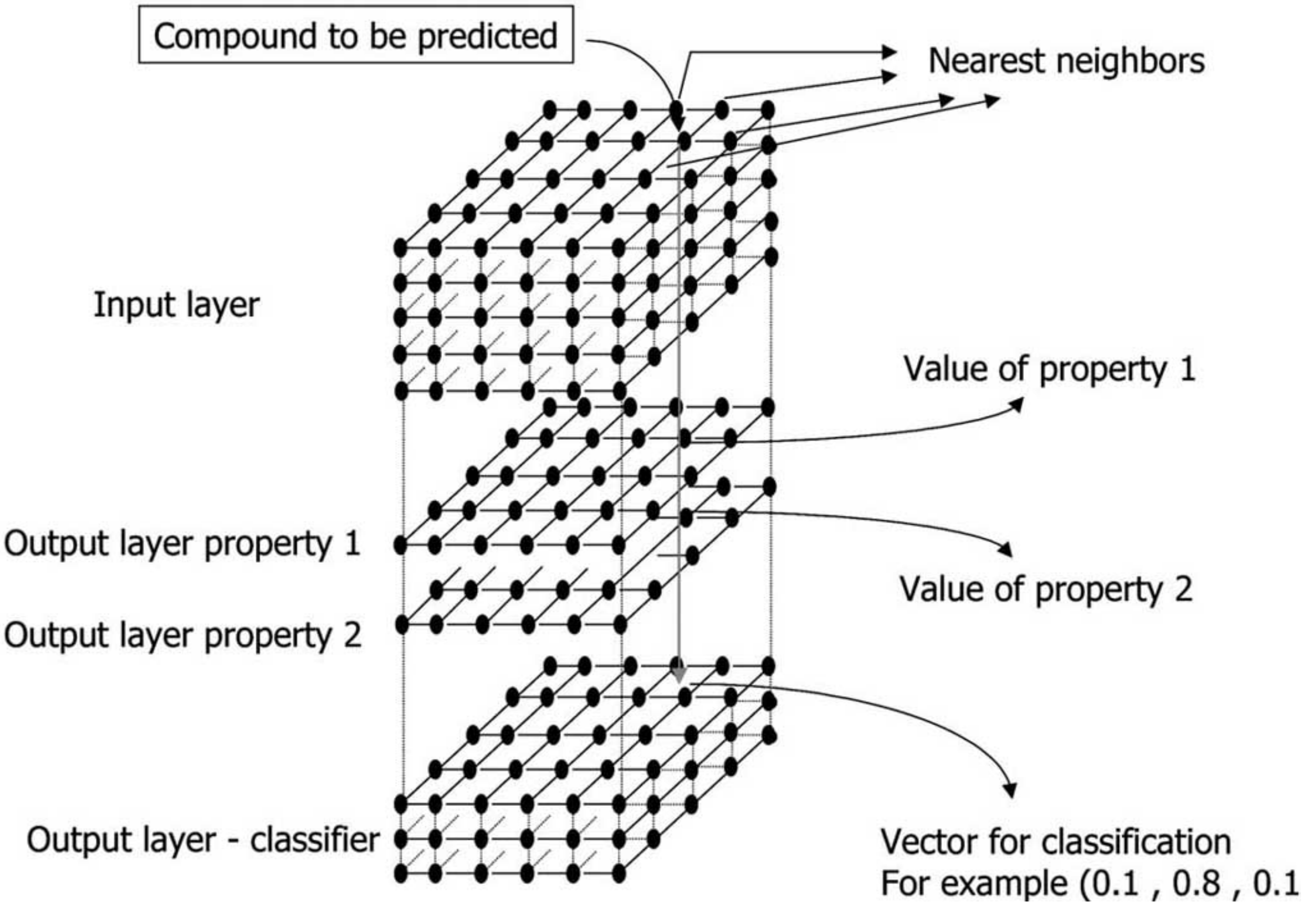
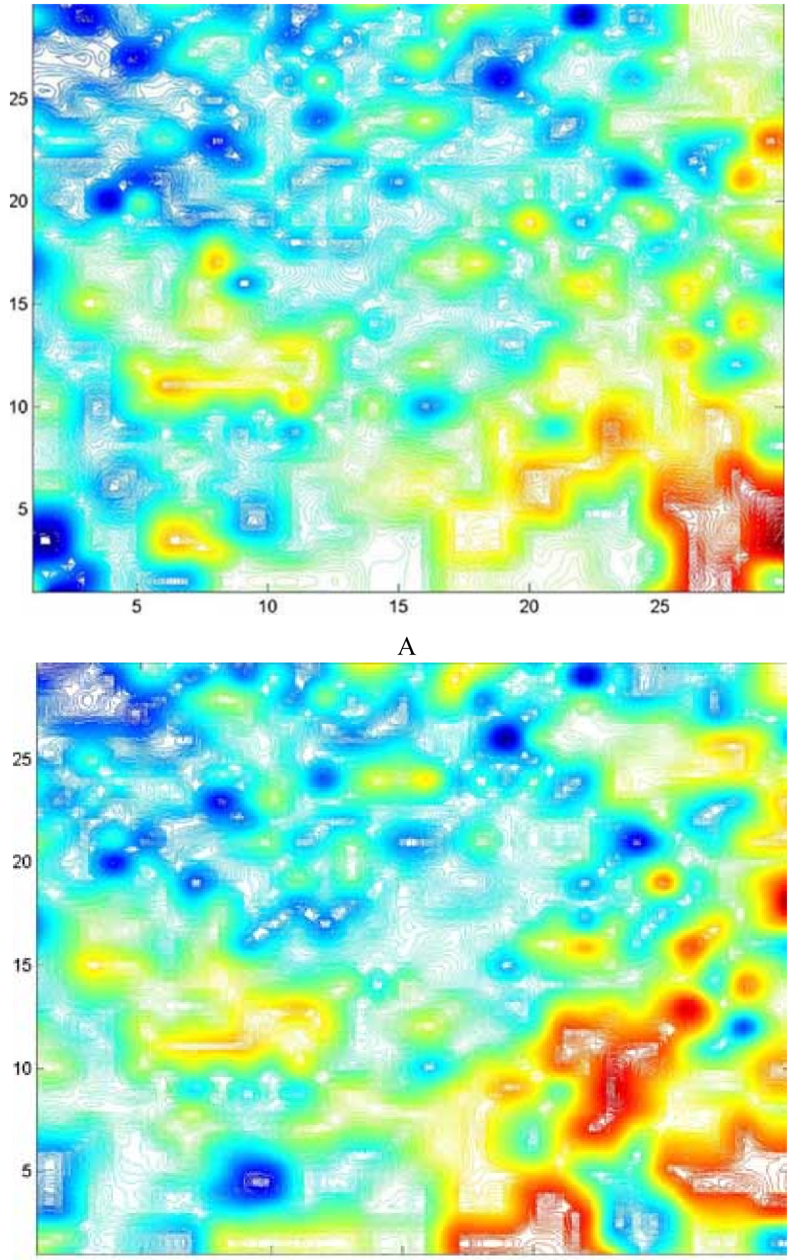


**Fig. (1).** Structure of CP NN.



A

B

**Fig. (2).** Distibution of water/octanol partition coefficient log P (figure A) and aquatic toxicity (figure B) for Duluth data set. Aquatic toxicity is defined as lethal concentration toward fish fathead minnow - Pimephales promelas. 562 compounds were described by 169 descriptors including log P. The map of dimension 30x30 was trained with 800 learning epochs.

analyzing individual descriptor layers in SOM one recognizes the importance of individual descriptors (see Fig. (**1**)). The comparison of contour plots of a descriptor layer and response surface shows the relationship between descriptor and property (see Fig. (**2**)).

In construction of SOM or CP NN models the following technical parameters must be determined, the minimal and maximal learning rates ($\eta_{min}$ and $\eta_{max}$), number of learning epochs and the dimension of the network. For the dimension of the network authors in references [14, 15] propose a rule of thumb, which states that the number of neurons should be one to three times the number of objects in training set. Alternatively, the dimension can be set as a larger dimension with all occupied neurons [16].

Usually in QSAR studies the number of descriptors overwhelms the number of objects. In SOM and CP NN there is formally no limit on dimension of representation vectors. However, one must be aware that in the algorithm all descriptors are treated equivocally and when a large number of descriptors is used an important descriptor can be diminished by a noise of others [17]. Large number of descriptors is used when molecular shapes [18], spectra [19], or other encoding of 3D molecular structures are taken to represent the molecules [20, 21].

The statistical parameters, which describe the quality of models, are mostly comparable for linear models and the SOM or CP NN models [22, 23]. It is to emphasize that in SOM and CP NN modeling one obtains more information about the data set from a single model [24]. In reference [25] authors report CP NN and multiple linear regression results for the set of 105 flavonoids studying the inhibitory activity of enzyme $p56^{lck}$. The results are in favor of CP NN method. The correlation coefficients for predicted test set are r = 0.71 for linear method and r = 0.82 for CP NN. Other examples of application of SOM in drug research are given in reference [10] and references therein.

## Clustering versus classification

Due to the basic property of the mapping to locate similar objects close to each other the SOM and CP NN are tools for clustering. Visual inspection of the map enables us to recognize the clusters and thus the similarity relationships within the data set [26]. The classification is conceptually different. It is an ordering of objects into exactly predefined classes. The CP NN with the strategy of 'multi-dimensional output layer' is a suitable tool for classification [21, 27, 28]. By ordering of objects into n-classes the n-dimensional output layer must be defined. Output variable (property) of a compound belonging to i-th class is described with n-dimensional vector with 'one' on the i-th position and 'zero' on all other positions. The learning runs as described above where n different response surfaces are created. In a prediction one obtains as a result n-dimensional vector with components expressed as real numbers between zero and one. Each number represents affiliation to a particular class. Two situations can occur. First, one component is essential larger than others. In such a case the predicted object unambiguity belongs to the pointed class. Second, more components are approximately the same. In such a case the object is set to several classes. In a very peculiar case all components are approximately equal. This means that the model can not decide in which class the element belongs, and this is a very valuable result. We know *a priori* that the model is not able to describe this particular object [21, 28, 29]. In reference [29] authors propose a model for classification of compounds on evidence of their endocrine disruption activity. The model was built with 106 compounds, which were selected from 553 chemicals inspected by the European Union Commission.

## Outliers

The SOM gives valuable information of similarity relationships among objects (molecules). Particularly interesting is the analysis of molecules that are located on the same neuron. The model recognizes such molecules as identical, or with the other words their representation vectors are too similar to be discriminated by the model. Furthermore, in the CP NN the output layer can be analyzed. It happens that two or more compounds with very different activities are located on the same neuron. We conclude that one is an outlier. To realize which one is the outlier we analyze the neighborhood of the neuron. If the neighboring neurons are occupied by active compounds the non active compound, is an outlier and vice versa. Existing of such outliers is not a disaster it solely means that our descriptors are not complete. In reference [22] structure-mutagenicity relationship was studied on the set of 95 aromatic and heteroaromatic compounds. The quality of linear and CP NN models is comparable. Beside the comparison of methods authors studied how different descriptors, which describe the molecular structures on different hierarchical levels, influence the selection of outliers.

## Descriptor Selection

This is an important issue in QSAR studies. Nowadays hundreds of descriptors can be easily calculated from molecular structures [6-8]. The SOM can be used for these purposes in different ways. The SOM can be trained with 'transposed matrix', i.e., the data matrix where the roles of objects and descriptors are exchanged. In the resulting SOM the descriptors are arranged according to the similarity relationship into clusters. One selects from each cluster representative descriptors [16, 29]. In reference [30] authors studied SOM of 169 descriptors plus toxicity on the set of 562 compounds of Duluth data set [31]. In the map three descriptors were found in the closest neighborhood of toxicity: log P, weight of the free compound, and the valence index of Kier and Hall. Fig. (**2**) shows the distribution of log P and the distribution of toxicity values over the map. In report [16] the SOM was applied for desriptor selection on set of 95 aromatic and heteroaromatic amines. From 240 descriptors a handful descriptors was systematically selected. Models built with four, five, or 36 descriptors have cross validation correlation coefficients r on the interval $0.83 < r < 0.85$.

Reference [32] reports the comparison between linear regression and CP NN results for 28 flavonoids (a subset of 105 flavonoids mentioned above) represented with multi-dimensional vectors, which encoded the 3D structures. Multiple linear regression was used to select the most relevant descriptors. In reference [23] authors studied mutagenic potency of 12 trimethylimidazopyridine isomers. In both last examples the linear method outperforms the CP NN.

## Testing and Validation

The testing of models is the crucial point in model developing. In reference [33] authors suggest that the final validation of models should be done with external validation set. However, in the model development several tests using the available data are recommended.

### *Recall Ability Test*

In this test the property values are predicted for the objects of training set. It shows how the model recognizes the training data. It is known that this test overestimates the quality of CP NN models. With CP NN models it is often possible to achieve a good fit to training set, but the model is poor in prediction. For objective estimation of the quality of models the leave-one-out and further test methods with independent sets must be taken.

### *Leave-One-Out Cross Validation*

Here the objects one after another are selected out, whereas for every selected object the model is built with remaining ones. In the next step this model is used to predict the property for the selected object. The test gives information on prediction ability and on the quality of data set. It is often used to set the technical parameters of model.

### *Division in Training and Test Set*

For testing of models we sometimes divide the data set into two subsets, training set is used to build the model, testing set to test it. To obtain reasonable predictions for the test set the training set must contain information of entire descriptor space. If the SOM is used for the division the Kohonen map is divided into sub-parcels selecting the objects for training and test set from each sub-parcel equivocally [34]. It is expected that such training set possesses the information content of the entire set [15, 35, 36]. The SOM is often used to divide the set into training and test set whereas latter other techniques are applied to build the models [15, 34, 37]. To prove how the both set cover the entire space is recommended to train the model with test set and test it with the training set. If the selection was good the both abilities are approximately equal. In some further tests the data can be scrambled, or random descriptors can be introduced [35-40]. In reference [36] authors studied the set of 95 aromatic amines with their mutagenic potency. After removal of two outliers the model was trained with 31 compounds and tested with 62 ones. The correlation coefficient for test set was found as r = 0.91 and when the roles of training set and test set were exchanged the correlation coefficient slightly decreased to r = 0.85. The model broke down when the output values were randomly permuted (r = 0.25). In reference [38] authors applied similar strategy in treatment of Duluth data set (see section above). For the test set the correlation coefficient was r = 0.77 and after exchanging of training and test set it was r = 0.75.

## CONCLUSIONS

We try to describe the properties and advantages of SOM and CP NN as tools in QSAR/QSPR modeling. A basic advantage is that different informations can be withdrawn from the single model. Analysis of clusters in network gives information on similarity relationship within the data set, the same model can be used for prediction of variables and for classification. Furthermore, the analysis of individual descriptor layers shows the role of descriptor in the model. On the other hand, the methods have some disadvantages. First, for the construction of reliable models some technical parameters must be set. We discussed the testing methods for the setting of parameters; however, a doubt on selection of proper parameters always exists. Second, SOM and CP NN do not extrapolate. Even the objects, which are completely out of domain are situated somewhere in SOM and CP NN gives a prediction, which is on the interval defined with the training data. Both shortcomings are avoided by carefully using of the methods. A scrutinizing of the input data and the results is necessary during the entire course of modeling.

## ACKNOWLEDGEMENTS

Author thanks Ministry of education, science and sport of Republic of Slovenia, which supported this work over decades under different contracts (034 0507, 034 0508, P1 017) and EU Commission, which supported this work under projects: COPERNICUS, IMAGETOX and Marie-Curie training site.

## ABBREVIATIONS

| | | |
|---|---|---|
| SOM | = | Self-organizing Map |
| CP NN | = | Counter Propagation Neural Network |
| QSPR/QSAR | = | Quantitative Structure-Property/Activity Relationship |
| TSCA | = | Toxic Substances Control Act |

## REFERENCES

[1] Auer, C.M.; Nabholz, J.V.; Baetcke, K.P. *Environ. Health Perspect.,* **1990**, *87*, 183.

[2] Allanou, R.; Hansen, B.G.; Bilt, Y.v.d. Public availability of data on EU high production volume chemicals. EUR 18996EN, © European Communities, **1999**.

[3] Hansch, C.; Leo, *A. Exploring QSAR. Fundamentals in chemistry and biology*. ACS professional reference book, American chemical society: Washington DC, **1995**.

[4] Mlinsek, G.; Novic, M.; Kotnik, M.; Solmajer, T. *J. Chem. Inf. Comput. Sci.*, **2004**, *44,* 1 (in press).

[5] Zuperl, S.; Mlinsek, G.; Novic, M.; Solmajer, T.; Zupan, J. In 11th International workshop on quantitative structure-activity relationships (QSAR) in the human health and environmental sciences [also] QSAR 2004 : 9-13 May, 2004, Liverpool, United Kingdom : abstract book., (P2-13), **2004**; p.37.

[6] Todeschini, R.; Consonni, V. In *The handbook of molecular descriptors. Series of methods and principles in medicinal chemistry*; R. Mannhold, H. Kubinyi, G. Timmerman, Eds.; Wiley: New York, **2000**; Vol. 11.

[7] Diudea, M.V. *QSPR/QSAR studies by molecular descriptors*, Nova Science Publishers, Inc.: Huntington, **2001**.

[8] Leardi, R. Nature-inspired methods in chemometrics: Genetic algorithms and artificial neural networks, Elsevier: Amsterdam, **2003**.

[9] Zupan, J.; Gasteiger, J. *Neural networks in chemistry and drug design*, Wiley-VCH: Weinheim, **1999**.

[10] Manallack, D.T.; Livingstone, D.J. *Eur. J. Med. Chem.,* **1999**, *34*, 195.

[11] Bienfait, B.; Gasteiger, J. *J. Mol. Graphics Mod.*, **1997**, *15*, 203.

[12] Kawakami, J.; Hoshi, K.; Ishiyama, A.; Miyagishima, S.; Sato, K. *Chem. Pharm. Bull.*, **2004**, *52*, 751.

[13] Hecht-Nielsen, R. *Appl. Optics,* **1987**, *26*, 4979.

[14] Chen, L.; Gasteiger, J. *J. Am. Chem. Soc.,* **1997**, *119*, 4033.

[15] Guha, R.; Serra, J.R.; Jurs, P.C. *J. Mol. Graphics Mod.,* **2004**, *23*, 1.

[16] Jezierska, A.; Vracko, M.; Basak, S.C. *Molecular Diversity*, **2004** (in press, available on-line).

[17] Kewley, R.H.; Embrechts, M.J.; Breneman, C. *IEEE T Neural Network*, **2000**, *11*, 668.

[18] Anzali, S.; Barnickel, G.; Krug, M.; Sadowski, J.; Wagener, M.; Gasteiger, J.; Polanski, J. *J. Comp. Aided Mol. Design.* **1996**, *10*, 521.

[19] Bursi, R.; Dao, T.; Wijk, T.V.; Gooyer, M.d.; Kellenbach, E.; Verwer, P.J. *Chem. Inf. Comput. Sci*., **1999**, *39*, 861.

[20] Vracko, M. *J. Chem. Inf. Comput. Sci*., **1997**, *37*, 1037.

[21] Vracko, M.; Novic, M.; Zupan, J. *Anal. Chim. Acta,* **1999**, *384*, 319.

[22] Vracko, M.; Mills, D.; Basak, S.C. Environ. Toxicol. Pharmacol., **2004**, *16*, 25.

[23] Vracko, M.; Szymoszek, A.; Barbieri, P. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 352.

[24] Bienfait, B. *J. Chem. Inf. Comput. Sci*., **1994**, *34*, 890.

[25] Novic, M.; Nikolovska-Coleska, Z.; Solmajer, T. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 990.

[26] Zupan, J.; Novic, M.; Li, X.; Gasteiger, J. *Anal. Chim. Acta*, **1994**, *292*, 219.

[27] Gini, G.; Testaguzza, V.; Benfenati, E.; Todeschini, R. *Chemometr. Intell. Lab. Syst.,* **1998**, *43*, 135.

[28] Vracko, M. *SAR and QSAR in Environ. Res.*, **2000**, *11*, 103.

[29] Roncaglioni, A.; Novic, M.; Vracko, M.; Benfenati, E. *J. Chem. Inf. Comput. Sci.,* **2004**, *44*, 300.

[30] Barbieri, P.; Piclin, N.; Szymoszek, A.; Novic, M.; Vracko, M.; Benfenati, E. In *Meeting of Slovenial Chemical Society* ***2001****: abstract book*, P. Glavic, D. Brodnjak-Voncina, Eds.; FKKT Maribor, **2001**; pp. 74-79.

[31] Russom, C.L.; Brandbury, S.P.; Broderius, S.J.; Hammermeister, D.E.; Drummond, D.A. *Environmental Toxicology and Chemistry*, **1997**, *16*, 948.

[32] Novic, M.; Vracko, M. *Chemometr. Intell. Lab. Syst.,* **2001**, *59*, 33.

[33] Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model*., **2002**, *20*, 269.

[34] Simon, V.; Gasteiger, J.; Zupan, J. *J. Am. Chem. Soc.*, **1993**, *115*, 9148.

[35] Vracko, M.; Gesteiger, J. *Internet Electronic Journal of Molecular Design,* **2002**, *1*, 527.

[36] Valkova, I.; Vracko, M.; Basak, S.C. *Anal. Chim. Acta*, **2004**, *509*, 179.

[37] Golbraikh, A.; Bernard, P.; Chretien, J.R. *Eur. J. Med. Chem*., **2000**, *35,* 123.

[38] Mazzatorta, P.; Vracko, M.; Jezierska, A.; Benfenati, E. *J. Chem. Inf. Comput. Sci.,* **2003**, *43*, 485.

[39] Embrechts, M.J.; Arciniegas, F.; Ozdemir, M.; Breneman, C.; Bennett, K. Data mining using 2-D neural network sensitivity analysis for molecules. ANNIE Conference, St. Luis, Missouri, November **2001**, ASME: **2001**.

[40] Arciniegas, F.; Bennett, K.; Breneman, C.; Embrechts, M.J. Molecular database mining using self-organizing maps for the design of novel pharmaceuticals. ANNIE Conference, St. Luis, Missouri, November **2001**, ASME: **2001**.

**Received: September 17, 2004** **Accepted: October 20, 2004**