Modeling Toxicity by Using Supervised Kohonen Neural Networks

Paolo Mazzatorta,*,†,‡ Marjan Vračko,‡ Aneta Jezierska,‡,§ and Emilio Benfenati†

Istituto Mario Negri, via Eritrea 62, 20157 Milan, Italy, National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, P.O. Box 3430, Slovenia, and University of Wroclaw, Faculty of Chemistry, F. Joliot-Curie 14, 50-385 Wroclaw, Poland

Received October 11, 2002

Counterprogation neural network is shown to be a powerful and suitable tool for the investigation of toxicity. This study mined a data set of 568 chemicals. Two hundred eighty-two objects were used as the training set and 286 as the test set. The final model developed presents high performances on the data set $R^2 = 0.83$ ($R^2 = 0.97$ on the training set, $R^2 = 0.59$ on the test set). This technique distinguishes itself also for the ability to give to the expert two-dimensional maps suitable for the study of the distribution/clustering of the data and the identification of outliers.

I. INTRODUCTION

Quantitative Structure−Activity Relationships (QSARs) are based on the assumption that the structure of a molecule must contain the features responsible for its physical, chemical, and biological properties and on the possibility of representing a molecule by numerical descriptors. In the past years these techniques were successfully applied to a wide variety of physical, chemical, biological, and technological properties.[1]

The use of QSAR for the determination of toxicity responds to the need to understand and predict the consequence of the thousands of chemicals, that every year are synthesized,[2] to human health, wildlife, and environment. The experimental determination of toxicity is facing several problems. The reliability of some experimental data is questionable, because many different laboratory conditions such as sex, age, and health of testing animals influence the results. The testing is expensive and of long duration. Public opinion demands a reduction of experiments on animals due to ethical reasons. The computational estimation of toxicity is also difficult.[3−6] The term "toxicity" defines just a biological end-point, including several different mechanisms at the molecular level. Data sets usually consist of compounds acting by different mechanisms making the use of linear models questionable. Artificial Neural Networks (ANN), such as a nonlinear method, seem to be more appropriate for the modeling of toxicity of data sets of diverse compounds.[7−13]

In the present study a set of 568 compounds was treated with the CounterPropagation Artificial Neural Network (CP ANN) method to build models for prediction of acute toxicity (LC_{50}). Section II describes the data set mined and the descriptor calculated for the characterization of the molecules; some computational details of the method and the software used are also presented. The models developed and their analysis are included in section III. The last section contains conclusions and suggestions for further works.

II. MATERIAL AND METHODS

Data Set. The U.S. Environmental Protection Agency[14−17] provided information to build up a data set, starting from a revision of experimental data from the literature, referred to as acute toxicity 96 h (LC50), for the fathead minnow (*Pimephales promelas*) expressed as log(mmol/L). The data set contains 568 organic compounds, commonly used in industrial processes. This is a large set of compounds belonging to different chemical classes: a positive characteristic is the homogeneity and reliability of this toxicological data. A preprocessing processing phase is necessary to make the data suitable for modeling. Thus, the toxicity value was normalized using the expression 1 (eq 1).

$$y_{new,i} = \frac{\log_{10}(LC_{50}) - \min(\log_{10}(LC_{50}))}{\max(\log_{10}(LC_{50})) - \min(\log_{10}(LC_{50}))} \quad (1)$$

Here y_i is the normalized value and $LC_{50,i}$ is the original value; $\min(\log_{10}(LC_{50}))$ and $\max(\log_{10}(LC_{50}))$ are the minimum and the maximum values, respectively.

Descriptors. Descriptors are used to mathematically characterize the molecules. A large number of descriptors was calculated by Istituto di Ricerche Farmacologiche "Mario Negri" using different software: Hyperchem 5.0 (Hypercube Inc., Gainsville, FL, U.S.A.), CODESSA 2.2.1 (SemiChem Inc., Shawnee, KS, U.S.A.), Pallas 2.1 (CompuDrug, Budapest, Hungary). Out of the hundreds of descriptors proposed by these software just 150 gave a nonconstant or nonmissing value for all the objects. The set of descriptors resulting can be split, according to the classification present in the software CODESSA,[19−21] into six categories: constitutional descriptors, depending on the number and type of atoms, bonds, and functional groups; geometrical descriptors, which give molecular surface area and volume, moments of inertia, shadow area, projections, and gravitational indices; topological descriptors, which are molecular connectivity indices, related to the degree of branching in the compounds;

* Corresponding author phone: +39-02-39014499; fax: +39-02-39001916; e-mail: mazzatorta@marionegri.it.
† Istituto Mario Negri.
‡ National Institute of Chemistry.
§ University of Wroclaw.

10.1021/ci0256182 CCC: $25.00
Published on Web 03/06/2003

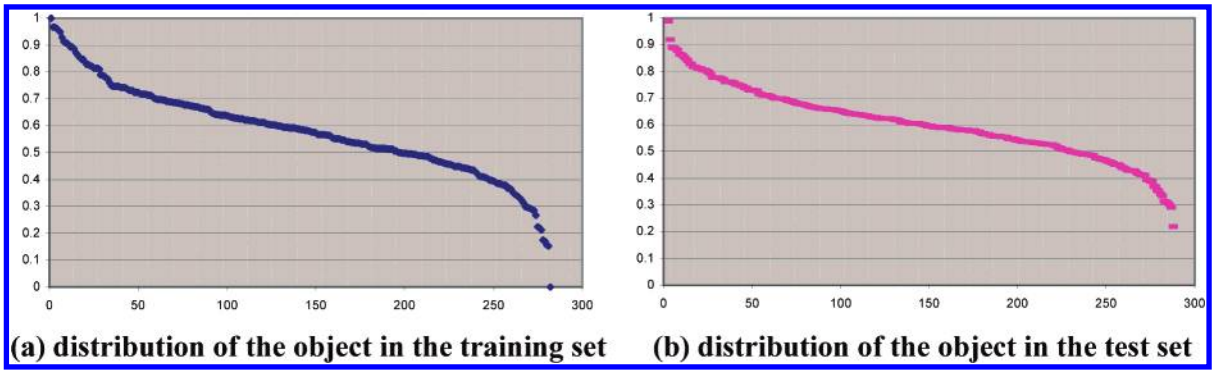

Figure 1. Distribution of the objects in the training (a) and the test sets (b). On the *x*-axis there is the number of compounds, on the *y*-axis there is the toxicity.

electrostatic descriptors, such as partial atomic charges and others depending on the possibility for some sites in the molecule to form hydrogen bonds; quantum-chemicals descriptors, i.e., total energy of the molecule, the energies of the lowest unoccupied and highest occupied orbital (HOMO and LUMO), ionization potentials, heat of formation, etc.; and physicochemical descriptors, such as logP. All the descriptors were normalized between zero and one using a range scaling procedure, maintaining the original distribution[18] (eq 2).

$$x_{\mathrm{new,i}} = \frac{x_{\mathrm{i}} - \min(x)}{\max(x) - \min(x)} \qquad (2)$$

SphereExcluder. For the selection of the training and the test set the SphereExcluder program,[22] developed at School of Pharmacy (University of North Carolina, U.S.A.), was used. The training and test sets must satisfy the following criteria: (i) representative points of the test set must be close to those of the training set; (ii) representative points of the training set must be close to representative points of the test set; and (iii) the training set must be diverse. This software divides the data set into training and test sets according to the following algorithm: (i) select a point and include it in the training set; (ii) build a sphere with radius R with the center in this point; (iii) all points within this sphere, except for the center, include in the test set; (iv) discard all points, which are within this sphere; and (v) if no points are left, stop, otherwise, go to step (i). The division is based on distances between points in the descriptor space and on random sphere center selection.

As result of this operation the data set was divided into training set (282 objects) and test set (286 objects) (Figure 1) such that $M_{\mathrm{test,train}} = 0$ (diversity index of the test set with respect to the training set) and $I_{\mathrm{train}} = 1$ (diversity of the training set), i.e., these indices take their optimal values.[23]

CP ANN. Hecht-Nielsen[24] and Dayhof[25] give a detailed description of CP ANN architecture and learning strategy. Architecture of the CP ANN is shown in Figure 2. CP ANN consists of two layers of neurons, input or Kohonen layer and output layer. The input or Kohonen layer gets input variables related to considered objects. During the learning, the target values (i.e. toxicity) are given to the output layer, which has the same topological arrangement of neurons as the Kohonen layer. Learning in the Kohonen layer is the same as in the Kohonen networks. This means a vector of input variables is presented to all neurons. The program

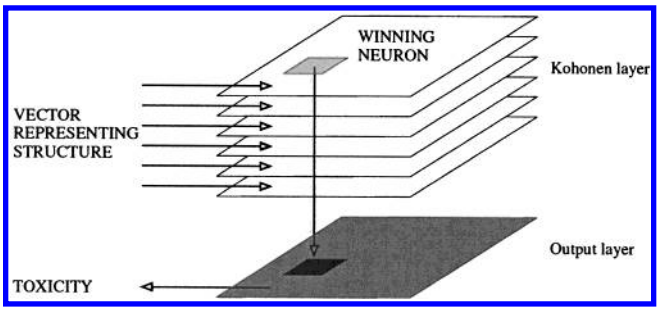

Figure 2. Architecture of CP ANN.

selects the neuron, which weights are closest to the input values. The chosen neuron is called the winning neuron. Two properties of the trained Kohonen layer are to emphasize. First similar objects are located close to each other, and, second, descriptors alone determine the learning procedure. The learning in output layer is different. The position of objects is projected to the output layer. In the next step, the weights in output layer are corrected in a way that they fit the output values (toxicities) of corresponding objects. Again, the output values (toxicities) do not influence the arrangement of an object in network.

A trained network provides some information on the data set. First, one can analyze the objects located on the same neuron. Such objects are recognized as identical. Second, a visual inspection of the map gives us information on clusters and similarity relationships between objects. Third, the trained network can be used for predictions. The prediction for a new object runs in two steps. First it will be situated in the Kohonen layer on the neuron with the most similar weights. This position is projected to the output layer, which provides the output value (toxicity).

III. MODELING

Selection and Testing of Models. Numerous models were tested on recall ability and with an external validation set to optimize the computational parameters.

In the recall ability test the model is used to predict the toxicities for training set. It is well-known that this test does not show the prediction ability of the models. Due to the architecture and learning strategy of CP ANN the recall ability of models is usually very high.[26]

The testing with the validation set is more reliable particularly when the validation set is selected independently from the modeling technique. Golbraikh and Tropsha[27] showed that to establish a reliable QSAR model the use of an external test set is necessary.

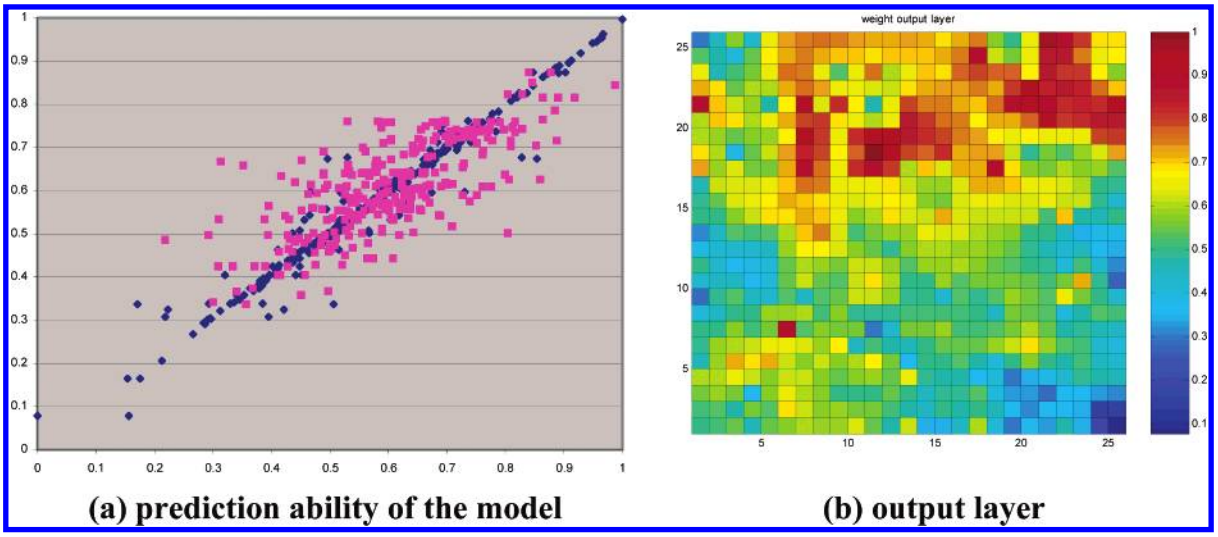

Figure 3. Results of the model. (a) On the *x*-axis there are the experimental values, on the *y*-axis there are the values predicted by the model; blue points represent the objects of the training set, pink points represent the object of the test set. (b) Output layer (blue represent high toxicity, red represent low toxicity).

Table 1. Setting Parameters for Model

network dimension			
nx	ny	epochs	R^2 test set
5	5	100	0.386
10	10	100	0.446
15	15	100	0.465
20	20	100	0.479
25	25	100	0.484
30	30	100	0.480
25	25	1000	0.527
25	25	2000	0.510

Table 2. Parameters Used for the Models

parameter	value	range
randomization of object sequence order	NO	yes; no
number of neurons in x direction (nx)	25	1−51
number of neurons in y direction (ny)	25	1−51
number of weights in each neuron	151	1−245
toroid boundary conditions	no	yes; no
type of neighborhood correction	triangular	flat; triangular; chef hat function; Mexican hat function
furthest neuron for corrections	25	1−75
maximal correction factor	0.50	0.1−0.9
minimal correction factor	0.01	0.00-maximal correction factor
epochs	1000	1-∞

Some results on tests are shown in Table 1. For this study the most influent parameters are the number of training epochs and the dimension of network. Other parameters only slightly influence the results. It is obvious from Table 1 that with increasing the number of epochs or increasing the dimension of network the correlation coefficient for validation set drops (overfitting effect).

The best models obtained on this data set were developed using the parameters shown in Table 2.

RESULTS AND DISCUSSION

Figure 3 shows an overview of the results obtained by the model developed. Figure 3a plots the values predicted by the model (*y*-axis) against the experimental value (*x*-axis) for the training set (blue points) and the test set (pink points). Figure 3b shows the output layer, the colorbar indicates the degree of similarity between neurons, and blue represents high toxicity and red low toxicity.

Table 3. Statistical Information of the Model: R^2 (Squared Correlation Coefficient), MAE (Mean Absolute Error), and MSE (Mean Squared Error)

training set		test set	
R^2	0.953	R^2	0.527
MAE	0.016	MAE	0.067
MSE	0.0013	MSE	0.0080

Some statistical information of the model developed is summarized in Table 3.

The model presents a high ability in predicting the toxicity of the training set that underlines the high skill of this tool in modeling. On the other hand the model has acceptable ability in predicting the test set too, which means that CP ANN are able to extract actual information and knowledge from the data set.

Determination of Outliers in Training Set. In the next step some of the objects were selected out as outliers. The aim to point the outliers is not to achieve better statistical parameters, but they are excluded because their toxic character cannot be described with other compounds in the set.

Two strategies of determination were applied. First, we analyzed the distances in the output layer (Figure 3b). This approach comes directly from the QSAR techniques, because elements with similar structure are close to each other in the space of the activity, i.e., our maps. For the evaluation of the level of dissimilarity ($ds_{nx,ny}$) between neurons we used an average of the difference between the weight of the neuron examined in the plane and its neighborhood (eq 3):

$$ds_{nx,ny} = \left| \frac{w_{nx-1,ny-1} + w_{nx,ny-1} + w_{nx+1,ny-1} + w_{nx-1,ny} + w_{nx-1,ny+1} + w_{nx,ny+1} + w_{nx+1,ny+1} - 8*w_{nx,ny}}{8} \right| \quad (3)$$

Here $ds_{nx,ny}$ is the level of dissimilarity of the neuron in position (nx,ny), and $w_{nx,ny}$ is the weight of the neuron in position (nx,ny).

Table 4. Outliers for the Training Set (ds > 0.25)

ID	name	ds	nx	ny
302	saccharin sodium salt hydrate	0.271	6	7

Table 5. Possible Outliers for the Training Set (AE > 0.1)

ID	name	absolute error	nx	ny
55	acetonitrile	0.181	10	21
190	chloroacetonitrile	0.178	10	21
285	diphenylamine	0.169	15	1
465	N-vinylcarbazole	0.167	15	1
533	1,3-diethyl-2-thiobarbituric acid	0.150	25	24
590	oxamyl #1	0.147	25	24
365	*tert*-butyl acetate	0.133	4	22
515	cyclohexyl acrylate	0.131	4	22
47	2,2′-methylene bis(3,4,6-trichlorophenol)	0.100	21	1

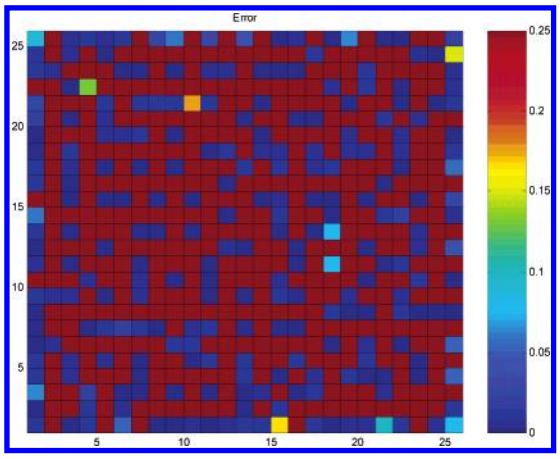

Figure 4. Map of the mean absolute error of each neuron for the training set. Neurons with an error of 0.25 (dark red on the map) are actually empty neurons (no object is associated with those neurons).

A level of dissimilarity > 0.25 was considered as an index of the presence of an outlier associated to that neuron. In this case (Table 4) the neuron (6,7) is the only one with a ds > 0.25 (ds = 0.27), and the object associated with this neuron, saccharin sodium salt hydrate (302), is therefore discarded as an outlier.

Second, we analyzed cases when two or more compounds are located on the same neuron. This means that the corresponding descriptors are too similar to be discriminated by the neural network. If the compounds on the same neuron have similar properties, we get another confirmation of our model. If the properties differ essentially, there is a conflict situation. One or more compounds are outliers. Objects of the training set with an absolute error (AE) higher than 0.1 were analyzed as possible outliers of the model. Those objects are listed in Table 5.

It is easily recognizable from Figure 4 that the position of these objects on the map correspond to the neurons with the higher mean absolute error. After the network is trained every object is associated with its winning neuron. The average of absolute error of all the objects associated with a given neuron is the mean error for this neuron.

This suggests that some of the objects that are associated with these neurons are actually outliers, and their presence on that neuron drive the model to an error. For the selection of the real outlier the near neurons were analyzed.

From this analysis four objects, listed in Table 6, were clearly recognized as outliers.

Table 6. Outliers for the Training Set

ID	name	absolute error	nx	ny
465	N-vinylcarbazole	0.167	15	1
533	1,3-diethyl-2-thiobarbituric acid	0.150	25	24
365	*tert*-butyl acetate	0.133	4	22
47	2,2′-methylene bis(3,4,6-trichlorophenol)	0.100	21	1

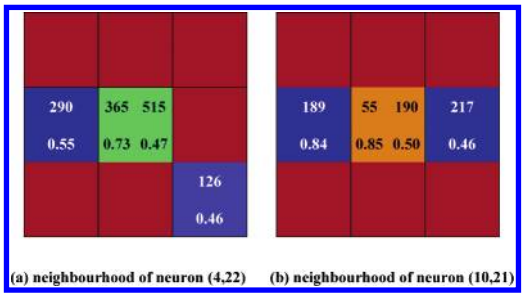

Figure 5. Neurons near neuron (4,22) (a) and neuron (10,21) (b). The objects and their toxicity value are shown for each neuron.

Table 7. Possible Outliers for the Test Set (AE > 0.2)

ID	name	absolute error	nx	ny
185	acrolein #1	0.353	9	19
550	1,3,5-trichloro-2,4-dinitrobenzene	0.306	6	3
18	salicylic acid Na+ #2	0.302	19	25
566	terbufos (counter)	0.268	23	4
73	2-butanone	0.233	8	21
377	isovaleraldehyde	0.229	7	22
462	α-bromo-2′,5′-dimethoxyacetophenone #1	0.224	16	9
525	2-decyn-1-ol	0.219	20	19
434	1-octyn-3-ol	0.217	16	16
401	α,α′-dichloro-p-xylene	0.205	9	3
170	4-chlorobenzaldehyde	0.200	6	2

For the objects associated with the neuron (10,21) it was not possible to determine an outlier because the near neurons do not present a clear trend. Thus, in this case, both the objects were kept.

An example of the procedure adopted for the selection of outliers is given in Figure 5. In Figure 5a (4,22) the identification of the outlier is possible because compound 365 has a toxicity value strongly different from the objects nearby; in Figure 5b the area nearby (10,21) does not give enough information to determine outliers.

Determination of Outliers in Test Set. The same procedure explained before was used to single outliers out in the test set. Because the worst ability of the model in predicting the object of the test set is foreseeable, only objects with an absolute error higher than 0.2 were analyzed. Those objects are listed in Table 7.

After the analysis of these objects on both the training set map (Figure 1) and the test set map (Figure 6) the following objects were recognized as outliers (Table 8).

Comparing the mean absolute error map of each neuron for the training set (Figure 4) with mean absolute error map of each neuron for the test set (Figure 6) is possible to notice that 2-butanone (73) is in the same area as acetonitrile (55) and chloroacetonitrile (190) and also in this case it is not possible to classify it as an outlier.

Training Again the Network. A new model was developed discarding from the sets the outliers selected in the previous steps. This model has the same parameters of the

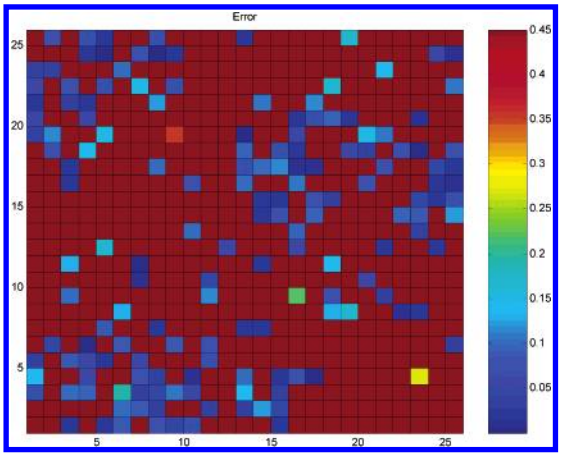

Figure 6. Map of the mean absolute error of each neuron for the test set. Neurons with an error of 0.45 (dark red on the map) are actually empty neurons (no object is associated with those neurons).

Table 8. Outliers of the Test Set

ID	name	absolute error	nx	ny
185	acrolein #1	0.353	9	19
550	1,3,5-trichloro-2,4-dinitrobenzene	0.306	6	3
18	salicylic acid Na+ #2	0.302	19	25
566	terbufos (counter)	0.268	23	4
377	isovaleraldehyde	0.229	7	22
462	α-bromo-2′,5′-dimethoxyacetophenone #1	0.224	16	9
525	2-decyn-1-ol	0.219	20	19
434	1-octyn-3-ol	0.217	16	16
401	α,α′-dichloro-p-xylene	0.205	9	3
170	4-chlorobenzaldehyde	0.200	6	2

previous one (see Table 2), and an overview of its performances is shown in Figure 7.

Statistical information of the model is summarized in Table 9.

The model shows a sensible improvement on both the training and the test sets. Discarding just 2.6% of the objects, the ability of the model in predicting the test set improves 7.6% for R^2 and 11% for MSE.

The analysis of possible outliers was conducted as it had been conducted previously for this model.

Table 9. Statistical Information of the Model: R^2 (Squared Correlation Coefficient), MAE (Mean Absolute Error), and MSE (Mean Squared Error)

training set		test set	
R^2	0.981	R^2	0.567
MAE	0.010	MAE	0.065
MSE	0.0005	MSE	0.0071

Table 10. Outliers for the Training Set (ds > 0.25)

ID	name	ds	nx	ny
45	methyl_sulfoxide	0.358	25	8
556	N,N-bis(2,2-diethoxyethyl)methylamine_#1	0.250	22	25

Table 11. Possible Outliers for the Training Set (AE > 0.1)

ID	name	absolute error	nx	ny
263	fensulfotion	0.124	18	1
430	carbophenotion	0.122	18	1

Table 12. Possible Outliers for the Test Set (AE > 0.2)

ID	name	absolute error	nx	ny
192	allyl alcohol	0.343	12	19
73	2-butanone	0.233	16	20
591	2,6-diisopropylaniline #1	0.225	23	10

The study of the map of the neurons weight for the output layer (Figure 7b) and the analysis of the ds of the neurons show the presence of outliers (Table 10). Possible outliers of this model according to their absolute error (Figure 8) are listed in Table 11 for the training set and in Table 12 for the test set.

The possible outliers for the training set belong to the same neuron, and from the analysis of the near neighborhood it is not possible to understand which one is the real outlier.

From the analysis of Figure 8a,b allyl alcohol (192) and 2,6-diisopropylaniline #1 (591) can be identified as outliers, but nothing can be deducted, also in this case, about 2-butanone (73).

Final Model. The outliers identified in the previous step were discarded, and then a new model (Figure 9) was developed training a neural network with the parameters listed in Table 2.

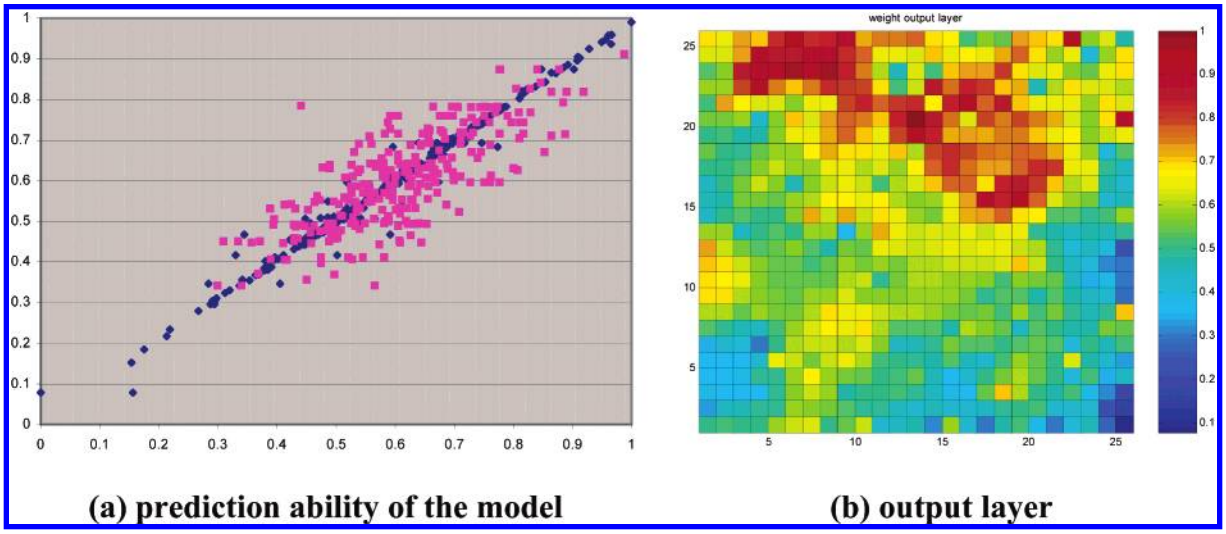

Figure 7. Results of the model. (a) On the *x*-axis there are the experimental values, on the *y*-axis there are the values predicted by the model; blue points represent the objects of the training set, pink points represent the object of the test set. (b) Output layer (blue represent high toxicity, red represent low toxicity).

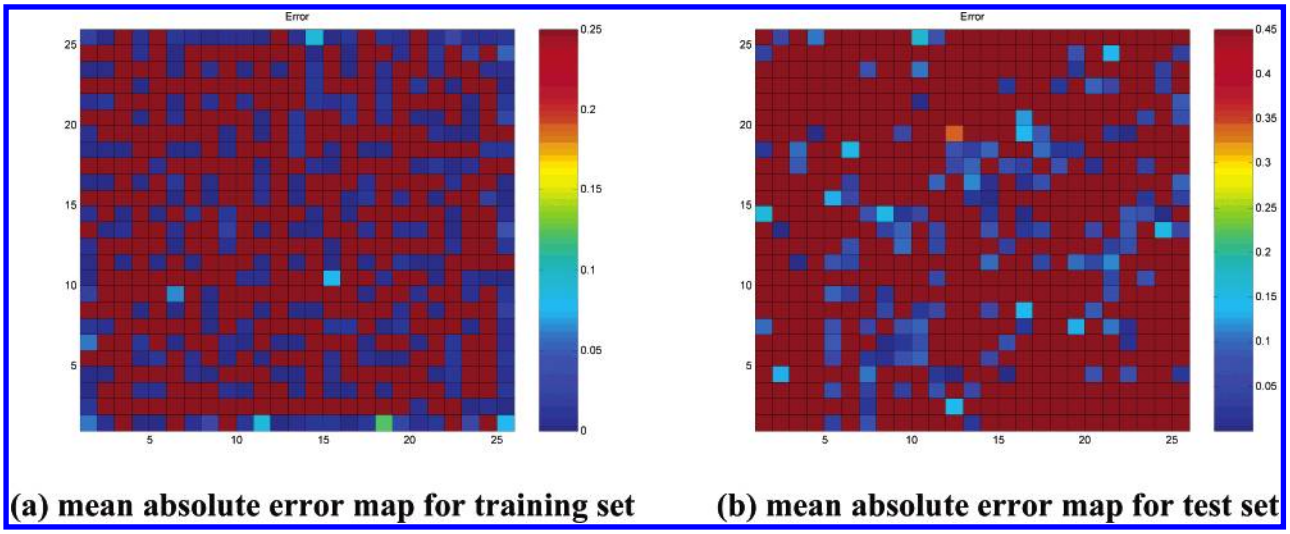

Figure 8. Map of the mean absolute error of each neuron for the training set (a) and for the test set (b). Empty neurons are represented by dark red (error = 0.25 for the training set (a) and error = 0.45 for the test set (b)).

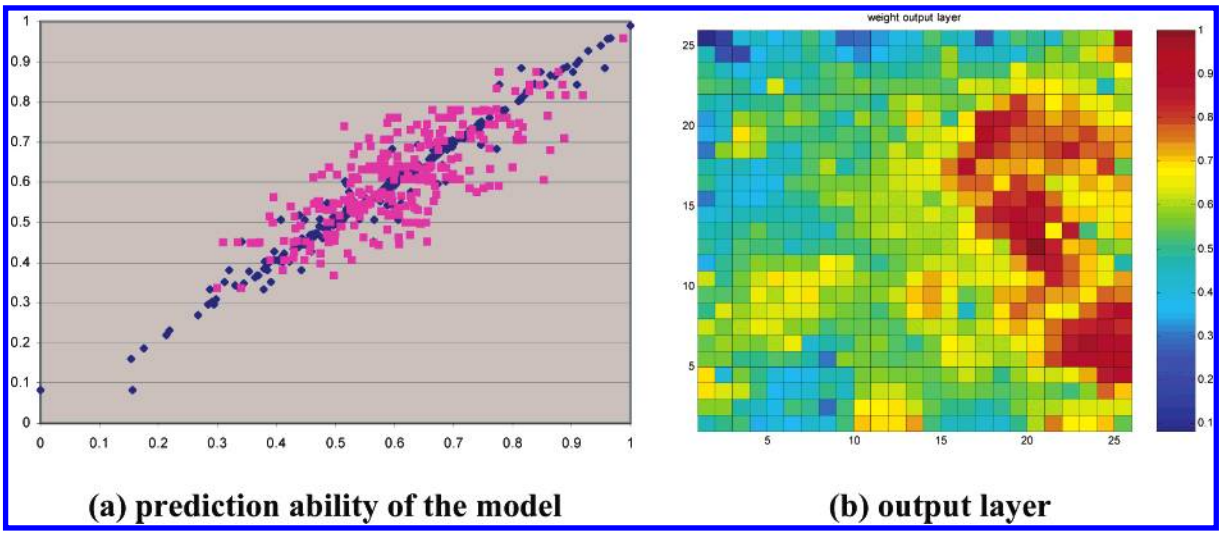

Figure 9. Results of the model. (a) On the *x*-axis there are the experimental values, on the *y*-axis there are the values predicted by the model; blue points represent the objects of the training set, pink points represent the object of the test set. (b) Output layer (blue represent high toxicity, red represent low toxicity).

Table 13. Statistical Information of the Model: R^2 (Squared Correlation Coefficient), MAE (Mean Absolute Error), and MSE (Mean Squared Error)

training set		test set	
R^2	0.971	R^2	0.592
MAE	0.014	MAE	0.064
MSE	0.0007	MSE	0.0065

Statistical information of the model are summarized in Table 13.

This model against a light worsening of the performances on the training set shows an improvement in predicting the test set, which is a more reliable parameter in evaluation of the model.

The study of the map of the neurons weight for the output layer (Figure 9b) and the analysis of the ds of the neurons do not show the presence of outliers. Possible outliers of this model according to their absolute error (Figure 10) are listed in Table 14 for the training set and in Table 15 for the test set.

Looking at Figure 10 it is not possible to establish the nature of these compounds because the compounds present in the near neurons were not enough to establish the trend of the area.

Table 14. Possible Outliers for the Training Set (AE > 0.1)

ID	name	absolute error	nx	ny
25	2-methyl-1,4-naphthoquinone	0.112	8	13
573	4-bromophenyl 3-pyridyl ketone	0.112	8	13

Table 15. Possible Outliers for the Test Set (AE > 0.2)

ID	name	absolute error	nx	ny
259	2-methyl-3-butyn-2-ol	0.247	18	12
193	2-propyn-1-ol #1	0.225	19	11
574	4-benzoylpyridine	0.203	8	13

Because no outliers were found this model can be considered as a final model.

The main statistics of prediction of the model developed are summarized in Table 16 for the whole data set (549 objects).

The robustness of the model was checked using the test set for training the network and the training set as a test.

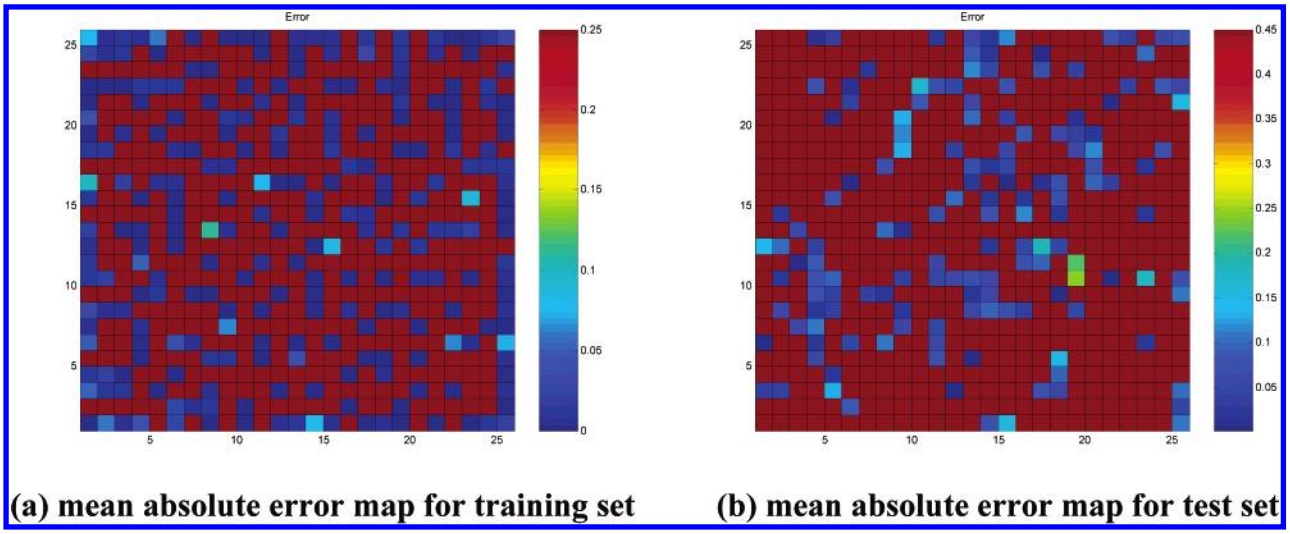

Figure 10. Map of the mean absolute error of each neuron for the training set (a) and for the test set (b). Empty neurons are represented by dark red (error = 0.25 for the training set (a) and error = 0.45 for the test set (b)).

Table 16. Main Statistics of Prediction: R^2 (Squared Correlation Coefficient), MAE (Mean Absolute Error), and MSE (Mean Squared Error)

	final model
R^2	0.826
MAE	0.0393
MSE	0.0055

The results obtained with the resulting model are comparable with the previous one ($R^2 = 0.97$ on training set, $R^2 = 0.56$ on test set).

IV. CONCLUSIONS

The application of the CP ANN method to the prediction of toxicity showed promising results. Due to the complexity and the variety of chemicals included in the data set mined the extraction of actual and reliable information is difficult, but the use of such a data set is fundamental in order to obtain general and robust models. The ability of CP ANN to cluster the compounds respecting input variables on a two-dimensional layer is a powerful tool for the analysis of the models developed and the identification of possible outliers.

The present study started from the analysis of 568 compounds, which were split into the training set (282) and the test set (286). From the model technique independent methods were applied to get the training and test set division. The division was carried out entirely in descriptors space, but the distribution of toxicity in both sets are quite similar. This method was shown to be very powerful in the split of the whole information of the data set. The model developed switching the two data sets generated had, in fact, comparable results. This property is essential for a suitable QSAR approach.

Preliminary models were developed to find out the presence of possible outliers. The final model, obtained after discarding 19 outliers (3%), showed encouraging results ($R^2 = 0.97$ on training set, $R^2 = 0.59$ on test set).

In the present study 150 descriptors were used to represent a compound, but it is believed that some descriptors play the most important role in these models.

ACKNOWLEDGMENT

The authors gratefully acknowledge Prof. A. R. Katritzky (Gainesville, Florida) and Prof. M. Karelson (Tartu, Estonia) for the use of CODESSA. Dr. A. Golbraikh and Dr. A. Tropsha (School of Pharmacy, University of North Carolina, U.S.A.) are kindly thanked for the use of SphereExluder. This work is partially funded by the EU under contract HPRN-CT-1999-00015.

REFERENCES AND NOTES

(1) http://clogp.pomona.edu/medchem/chem/qsar-db/.
(2) Chemical Abstracts Service, ACS, 2540 Olentangy River Road, P.O. Box 3012, Columbus, OH 43210.
(3) Cronin, M. T. D.; Dearden, J. C. QSAR in Toxicology 1. Prediction of Aquatic Toxicity. *Quant. Struct.-Act. Relat.* **1995**, *14*, 1−5.
(4) Cronin, M. T. D.; Dearden J. C. QSAR in Toxicology 2. Prediction of Acute Mammalian Toxicity and Interspecies Relationships. *Quant. Struct.-Act. Relat.* **1995**, *14*, 117−120.
(5) Cronin, M. T. D.; Dearden J. C. QSAR in Toxicology 3. Prediction of Chronic Toxicities. *Quant. Struct.-Act. Relat.* **1995**, *14*, 329−334.
(6) Cronin, M. T. D.; Dearden J. C. QSAR in Toxicology 4. Prediction of Nonlethal mammalian toxicological endpoints and expert systems for toxicity prediction. *Quant. Struct.-Act. Relat.* **1995**, *14*, 518−523.
(7) Benfenati, E.; Gini, G. Computational predictive programs (expert systems) in toxicology. *Toxicology* **1997**, *119*, 213−225.
(8) Gini, G.; Katritzky, A. R.. *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools;* AAAI Press: Menlo Park, CA, 1999; pp 40−43.
(9) Gini, G.; Lorenzini, M.; Benfenati, E.; Grasso, P.; Bruschi, M. Predictive Carcinogenicity: A Model for Aromatic Compounds, with Nitrogen-Containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1076−1080.
(10) Neagu, C.-D.; Aptula, A. O.; Gini, G. Neural and Neuro-Fuzzy Models of Toxic Action of Phenols. *IEEE International Symposium 'Intelligent Systems' Methodology, Models, Applications in Emerging Technologies IS2002*; Sept 10−12, 2002; Varna, Bulgaria.
(11) Neagu, C.-D.; Benfenati, E.; Gini, G.; Mazzatorta, P.; Roncaglioni, A. Neuro-Fuzzy Knowledge Representation for Toxicity Prediction of Organic Compounds, *15th European Conference on Artificial Intelligence*, July 21−26 2002; Lyon, France.
(12) Vracko, M. A study of Structure-Carcinogenic Potency Relationship with Artificial Neural Networks. The Using of Descriptors Related to Geometrical and Electronic Structures. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1037−1043.
(13) Vracko, M.; Novic, M.; Zupan, J. Study of structure−activity relationship by a counterpropagation neural network. *Anal. Chim. Acta* **1999**, *384*, 319−332.
(14) ECOTOX, *ECOTOXicology Database System, Code List*. Prepared for U.S. Environmental Protection Agency, Office of Research, Laboratory Mid-Continent Division (MED), Duluth, Minnesota. By OAO Corporation Duluth Minnesota, February 2000.

(15) ECOTOX, *ECOTOXicology Database System, Data Field Definition*. Prepared for U.S. Environmental Protection Agency, Office of Research, Laboratory Mid-Continent Division (MED), Duluth, Minnesota. By OAO Corporation Duluth Minnesota, February 2000.
(16) ECOTOX, *ECOTOXicology Database System, User Guide*. Prepared for U.S. Environmental Protection Agency, Office of Research, Laboratory Mid-Continent Division (MED), Duluth, Minnesota. By OAO Corporation Duluth Minnesota, February 2000.
(17) Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Hammermeister, D. E.; Drummond, S. J. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (pimephales promelas). *Environ. Toxicol. Chem.* **1997**, *16*, 948−967.
(18) Mazzatorta, P.; Benfenati, E.; Neagu, D.; Gini, G. The Importance of Scaling in Data Mining for Toxicity Prediction. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1250−1255.
(19) Karelson, M.; Maran, U.; Wang, Y. L.; Katritzky, A. R. QSPR and QSAR models derived using large molecular descriptors spaces. A review of CODESSA applications. *Collect. Czech. Chem. C* **1999**, *64*, 1551−1571.
(20) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA Comprehensive Descriptors for Structural and Statistical Analysis*; Reference Manual, version 2.0, Gainesville, FL, 1994.
(21) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; WILEY−VCH: Mannheim, GER, 2000.
(22) Golbraikh, A., Trophsa A. Predictive QSAR Modeling based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. *J. Comput-Aided Mol. Design*. In press.
(23) Golbraikh, A. Molecular Dataset Diversity Indices and Their Application to Comparison of Chemical Databases and QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 414−425.
(24) Hecht-Neilson, R. Counter propagation Networks. *Appl. Optics* **1987**, *26*, 4979−4984.
(25) Dayhof, J. In *Neural Network Architectures, An Introduction*; Van Nostrand Reinhold: New York, 1990; p 192.
(26) Novic, M.; Vracko, M. Comparison of spectrum-like representation of 3D chemical structure with other representations when used for modelling biological activity. *Chem. Intell. Lab. Sys.* **2001**, *59*, 33−44.
(27) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.

CI0256182