An Integrated SOM-Fuzzy ARTMAP Neural System for the Evaluation of Toxicity

G. Espinosa,[†] A. Arenas,[#] and Francesc Giralt*[,†]

Departament d'Enginyeria Química and Departament d'Enginyeria Informàtica i Matemàtiques, Escola Tècnica Superior d'Enginyeria Química (ETSEQ), Universitat Rovira i Virgili, Av. dels Països Catalans, 26, 43007 Tarragona, Catalunya, Spain

Received June 25, 2001

Self-organized maps (SOM) have been applied to analyze the similarities of chemical compounds and to select from a given pool of descriptors the smallest and more relevant subset needed to build robust QSAR models based on fuzzy ARTMAP. First, the category maps for each molecular descriptor and for the target activity variable were created with SOM and then classified on the basis of topology and nonlinear distribution. The best subset of descriptors was obtained by choosing from each cluster the index with the highest correlation with the target variable and then in order of decreasing correlation. This process was terminated when a dissimilarity measure increased, indicating that the inclusion of more molecular indices would not add supplementary information. The optimal subset of descriptors was used as input to a fuzzy ARTMAP architecture modified to effect predictive capabilities. The performance of the integrated SOM-fuzzy ARTMAP approach was evaluated with the prediction of the acute toxicity LC_{50} of a homogeneous set of 69 benzene derivatives in the fathead minnow and the oral rat toxicity LD_{50} of a heterogeneous set of 155 organic compounds. The proposed methodology minimized the problem of misclassification of similar compounds and significantly enhanced the predictive capabilities of a properly trained fuzzy ARTMAP network.

INTRODUCTION

The design of new drugs based on the forecast of activity from molecular information only is a major challenge in pharmacology and chemistry. The traditional methods that incorporate synthesis in the design process are very reliable but usually laborious and expensive. On the other hand, computed assisted methods are time-consuming and very sensitive to the molecular input information selected. The main problem arises when trying to decide which molecular descriptors and algorithm should be used to build the computational model, i.e., the Quantitative Structure Activity Relationships (QSAR).

QSAR methods assume that the properties of chemical compounds, which are implicit in their molecular structure, can be established with a set of descriptors of reasonable dimension. This implies that the chemical properties of *similar* compounds are related. But, which is the meaning of the term *similar*? Some studies state that compounds are similar when they have the same action mechanism versus different physical, chemical, or biological conditions but not necessarily a comparable chemical structure. Sometimes the contrary is stated. Previous QSAR studies use both definitions for *similar* at their advantage. However, the determination of the action mechanism is usually very difficult, while it is usually easier to establish chemical similarities. For this reason, the majority of QSAR models are developed for sets of homogeneous compounds (families), based on the premise that they will have the same action mechanisms.

Models are built by establishing the relationships between known experimental properties and molecular features quantified by molecular descriptors, which typically include electronic information (e.g., the dipole moment) and/or measures of molecular shape (connectivity indices,[1−3] the Wiener index,[4] etc.). Recently, quantum information has been incorporated into QSPR/QSAR to better explain the property or activity of homogeneous and heterogeneous chemical compounds. Since the number of descriptors could be very large, statistical prescreening techniques are commonly used to select the most appropriate ones.[5,6]

The reduction of the number of descriptors usually involves the following steps:

(i) Exclusion of all descriptors that contribute with up to 90% of information already accounted for by other molecular indices;

(ii) Selection of only one descriptor in pairs with cross-correlations greater than 0.95;

(iii) Selection of the first descriptors when ranked following an orthogonalization procedure, using for example vector space descriptor analysis (VSDA).[5,6]

Self-organizing Kohonen feature maps (SOM)[7−10] could be an alternative to statistical prescreening techniques since they have been successfully applied to cluster molecules into three-dimensional predefined self-organized maps. The analysis of the shape and surface properties of those maps has provided valuable information about the biological activity of the molecules. These applications included a one-to-one mapping of a molecule into a single Kohonen network.[11−14] Thus, SOM can be an alternative for the selection of the best set of molecular descriptors needed to establish sound QSPR/QSAR models in difficult problems, such as in toxicity prediction. These models can be determined using either

* Corresponding author phone: +34-977-559638; fax: +34-977-558205; e-mail: fgiralt@etseq.urv.es.
[†] Departament d'Enginyeria Química.
[#] Departament d'Enginyeria Informàtica i Matemàtiques.

10.1021/ci010329j CCC: $22.00
Published on Web 01/26/2002

classical algorithms, such as partial least-squares or multilineal regression analysis, or neural networks.

The most widely applied neural network to build QSAR/QSPR models is back-propagation. Nevertheless, this neural system has some problems inherent to its architecture in relation to overtraining, overfitting, and network optimization. An alternative to improve both predictive capabilities and to establish a more transparent QSAR/QSPR methodology is the application of cognitive classifiers. Espinosa et al.[15,16] developed fuzzy ARTMAP based QSPRs models to estimate boiling points and critical properties of heterogeneous sets of organic compounds. The models obtained were superior to QSPRs obtained with optimized back-propagation architectures and to traditional group contribution methods reported in the literature. Fuzzy ARTMAP based models are an alternative to standard predictive algorithms and possess a number of distinctive features that overcome some of the limitations of back-propagation (feed-forward) neural networks. The most important of them are (i) continuous (online), fast and stable learning, and (ii) the ability to learn novel inputs and infrequent events without forgetting previously learned information by creating new input categories and output classes dynamically.

The purpose of the current work is to apply SOM to select key molecular information from any given pool of descriptors to build efficient QSAR models based on fuzzy ARTMAP. The methodology for the selection of descriptors uses SOM to establish the relative differences between molecules when diverse molecular information is used to describe them. A simple dissimilarity measure provides sufficient quantitative information to support the visual information on how different maps represent the intrinsic relations between the molecular structures.[17,18] The performance of the integrated methodology is illustrated with the development of two new fuzzy ARTMAP based QSAR for toxicity assessment: the first one for the acute toxicity, measured as lethal-dose LC_{50}, of 69 benzene derivatives[19,20] and the second one for the oral rat toxicity LD_{50} of a heterogeneous set of 155 organic compounds.

The current approach is related to the variable selection methods recently published by Zhen and Tropsha[21] and Ivanciuc et al.[22] The former study is based on the k-nearest-neighbor principle and selects the optimal subset of descriptors by using simulated annealing as a stochastic optimization algorithm. The latter study measures chemical diversity with quasi-orthogonal basis sets. Kohonen neural networks are applied in the current study to select first the molecular descriptors that best represent the diversity of all molecular information in relation to the target toxicity variables, while covariances are used afterward to include the more relevant information. The description of data sets and of molecular descriptors is included in the next section, which is followed by a description of the integrated SOM-fuzzy ARTMAP methodology. Finally, the results obtained for the LC_{50} of benzene derivatives and for the LD_{50} of organic compounds are presented and discussed.

Data Sets and Molecular Descriptors. Two different sets of toxicity data have been chosen to test the performance of the proposed integrated methodology for descriptor selection and QSAR model building. The first data set contains the acute toxicity (LC_{50}) of 69 benzene derivatives in the fathead minnow. Hall et al.[19] first studied this data set, which was used later by Gute and Basak[20] to compare them. The benzene substituents are amino, bromo, chloro, hydroxyl, methyl, methoxyl, and nitro groups. The complete list of compounds and their corresponding experimental toxicity values are shown in Table 1. This table also includes an overview of the best set of descriptors used to build the fuzzy ARTMAP-based QSAR for LC_{50}, which is expressed as the negative logarithm of the lethal concentration, $-\log(LC_{50})$, at which 50% of the exposed individuals die. The experimental values range from 3.07 to 6.37 log units.

The second data set includes 155 diverse organic compounds. The functional groups considered include alcohols, ketones, esters, carboxylic acids, aldehydes, ethers, nitriles, amines, and aromatic derivatives. The experimental values were collected from many literature sources. The complete list of compounds and their corresponding experimental acute toxicity values (LD_{50}) are included in Table 2, together with the best set of molecular descriptors used in the fuzzy ARTMAP model. The toxicity is expressed as the logarithm of the lethal dose, $\log(LD_{50})$, in mg/kg body wt/day, at which 50% of the individuals die. The experimental values range from 1.59 to 4.71 log units.

Both topological and quantum descriptors were included in the pool of molecular information. The topology of chemicals was accounted for by the following indices: (i) the connectivity indices from zero to four order ($^0\chi$, $^1\chi$, $^2\chi$, $^3\chi$, $^4\chi$); (ii) the hydrogen bonding and electronic contributions, represented by the Hansen hydrogen bond index and the polarity index, respectively, both estimated by fragment constant additions (the fragment values were determined from Hansen's work[23]); and (iii) the sum of atomic numbers and the kappa index. The first two groups of descriptors were generated using Molecular Modeling Pro 3.1[24] and were independent of the geometry optimization scheme. Semiempirical, PM3 Hamiltonian, geometry optimizations, and conformational searches for all the structures were carried out with MOPAC 6.0,[25] because of the relatively short computational times required, compared to ab initio calculations, and the availability of parametrization for a variety of atoms.

The following quantum chemical descriptors derived form semiempirical calculations were considered to describe molecular interactions: (i) the average molecular polarizability, which is related to inductive interactions in the molecule and measures the capacity to accept electrons; (ii) the dipole moment as a measure of the global polarity of the molecule; (iii) the number of doubly occupied (filled) MO levels; (iv) the electron−electron repulsion energy; (v) the electron−nuclear attraction energy; (vi) the resonance energy or differential between localized and delocalized π electrons in double bonds; and (vii) the exchange energy to account for the interaction involving two electrons. The sum of total one-center energies (electron−electron repulsion and electron−nuclear attraction) and the two-center terms (resonance and exchange energies) yield the total energy.

INTEGRATED SOM-FUZZY ARTMAP METHODOLOGY

Kohonen Self-Organizing Maps (SOM). Kohonen self-organizing maps constitute the neural system proposed in the present study to select the molecular features that best describe a given activity or property from a pool of molecular

Table 1. Best Set of Molecular Descriptors and Experimental Acute Toxicity LC_{50} for 69 Benzene Derivatives[a]

	information												
compd	ID	formula	$^0\chi$	$^1\chi$	$^2\chi$	$^3\chi$	$^4\chi$	N	NFL	AP	NNR	ENA	exp $-\log(LC_{50})$
2,3,4,5,6-pentachlorophenol	tr	C 6 H 1 O 1 Cl 5	1.01E+01	4.55E+00	3.80E+00	3.44E+00	1.72E+00	1.36E+02	3.30E+01	1.02E+02	8.59E+03	−1.72E+04	6.06E+00
1,2,3,4-tetrachlorobenzene	tr	C 6 H 2 Cl 4	9.68E+00	3.93E+00	3.22E+00	2.69E+00	1.24E+00	1.06E+02	2.70E+01	8.57E+01	5.97E+03	−1.20E+04	5.43E+00
2,3,4,5-tetrachlorophenol	tr	C 6 H 2 O 1 Cl 4	1.01E+01	4.07E+00	3.36E+00	2.82E+00	1.35E+00	1.28E+02	3.00E+01	9.09E+01	7.26E+03	−1.46E+04	5.72E+00
1,2,3-trichlorobenzene	tr	C 6 H 3 Cl 3	9.63E+00	3.44E+00	2.73E+00	2.13E+00	1.00E+00	9.00E+01	2.40E+01	7.45E+01	4.90E+03	−9.84E+03	4.89E+00
1,2,4-trichlorobenzene	tr	C 6 H 3 Cl 3	9.63E+00	3.44E+00	2.81E+00	1.85E+00	9.95E−01	9.00E+01	2.40E+01	7.60E+01	4.82E+03	−9.68E+03	5.00E+00
1,3,5-trichlorobenzene	tr	C 6 H 3 Cl 3	9.63E+00	3.43E+00	2.89E+00	1.48E+00	1.39E+00	9.00E+01	2.40E+01	7.65E+01	4.76E+03	−9.57E+03	4.74E+00
1,3,5-trinitrobenzene	tr	C 6 H 3 N 3 O 6	9.82E+00	3.22E+00	2.24E+00	1.36E+00	7.87E−01	1.08E+02	3.90E+01	8.77E+01	1.16E+04	−2.35E+04	5.29E+00
2,4,6-tribromophenol	tr	C 6 H 3 O 1 Br 3	1.25E+01	4.82E+00	4.34E+00	2.65E+00	2.79E+00	1.55E+02	2.70E+01	8.86E+01	5.88E+03	−1.18E+04	4.70E+00
1,2-dichlorobenzene	tr	C 6 H 4 Cl 2	9.57E+00	2.96E+00	2.23E+00	1.58E+00	7.10E−01	7.40E+01	2.10E+01	6.45E+01	4.01E+03	−7.85E+03	4.40E+00
1,4-dichlorobenzene	tr	C 6 H 4 Cl 2	9.57E+00	2.95E+00	2.31E+00	1.31E+00	6.81E−01	7.40E+01	2.10E+01	6.60E+01	3.94E+03	−7.70E+03	4.62E+00
1,2-dinitrobenzene	tr	C 6 H 4 N 2 O 4	9.70E+00	2.82E+00	1.87E+00	1.16E+00	6.35E−01	8.60E+01	3.10E+01	7.40E+01	8.26E+03	−1.67E+04	5.45E+00
1,3-dinitrobenzene	tr	C 6 H 4 N 2 O 4	9.70E+00	2.82E+00	1.87E+00	1.14E+00	6.52E−01	8.60E+01	3.10E+01	7.39E+01	7.87E+03	−1.60E+04	4.38E+00
1,4-dinitrobenzene	tr	C 6 H 4 N 2 O 4	9.70E+00	2.82E+00	1.87E+00	1.16E+00	6.35E−01	8.60E+01	3.10E+01	7.51E+01	7.78E+03	−1.58E+04	5.22E+00
2,4-dinitrophenol	tr	C 6 H 4 N 2 O 5	1.01E+01	2.96E+00	2.03E+00	1.25E+00	7.01E−01	9.40E+01	3.40E+01	8.02E+01	9.35E+03	−1.89E+04	4.04E+00
2,4-dichlorophenol	tr	C 6 H 4 O 1 Cl 2	9.94E+00	3.09E+00	2.44E+00	1.45E+00	8.89E−01	8.20E+01	2.40E+01	7.03E+01	4.93E+03	−9.94E+03	4.30E+00
bromobenzene	tr	C 6 H 5 Br 1	1.03E+01	2.89E+00	2.21E+00	1.26E+00	7.19E−01	7.60E+01	1.80E+01	5.76E+01	2.97E+03	−5.98E+03	3.89E+00
chlorobenzene	tr	C 6 H 5 Cl 1	9.52E+00	2.48E+00	1.73E+00	9.85E−01	5.60E−01	5.80E+01	1.80E+01	5.51E+01	3.11E+03	−6.06E+03	3.77E+00
4-nitrophenol	tr	C 6 H 5 N 1 O 3	9.95E+00	2.54E+00	1.69E+00	1.00E+00	1.95E+00	7.20E+01	2.60E+01	6.64E+01	5.84E+03	−1.18E+04	3.36E+00
2-chlophenol	tr	C 6 H 5 O 1 Cl 1	9.89E+00	2.62E+00	1.86E+00	1.17E+00	5.87E−01	6.60E+01	2.10E+01	5.97E+01	4.02E+03	−8.10E+03	4.02E+00
benzene	tr	C 6 H 6	9.46E+00	2.00E+00	1.15E+00	6.67E−01	3.85E−01	4.20E+01	1.50E+01	4.56E+01	2.37E+03	−4.58E+03	3.40E+00
4-hydroxy-3-nitroaniline	tr	C 6 H 6 N 2 O 3	1.15E+01	2.75E+00	1.93E+00	1.14E+00	6.47E−01	8.00E+01	2.90E+01	7.70E+01	7.18E+03	−1.45E+04	3.65E+00
phenol	tr	C 6 H 6 O 1	9.83E+00	2.13E+00	1.34E+00	7.56E−01	4.28E−01	5.00E+01	1.80E+01	5.08E+01	3.13E+03	−6.30E+03	3.51E+00
1,3-hydroxybenzene	tr	C 6 H 6 O 2	1.02E+01	2.27E+00	1.52E+00	8.30E−01	4.93E−01	5.80E+01	2.10E+01	5.57E+01	4.08E+03	−8.23E+03	3.04E+00
2,3,6-trinitrotoluene	tr	C 7 H 5 N 3 O 6	1.27E+01	3.65E+00	2.62E+00	1.76E+00	1.10E+00	9.20E+01	4.20E+01	9.82E+01	1.37E+04	−2.77E+04	6.37E+00
2,4,6-trinitrotoluene	tr	C 7 H 5 N 3 O 6	1.27E+01	3.65E+00	2.65E+00	1.76E+00	1.02E+00	9.20E+01	4.20E+01	9.74E+01	1.34E+04	−2.71E+04	4.88E+00
3,4-dichlorotoluene	tr	C 7 H 6 Cl 2	1.25E+01	3.37E+00	2.73E+00	1.81E+00	9.53E−01	8.20E+01	2.40E+01	7.37E+01	4.92E+03	−9.89E+03	4.74E+00
2,3-dinitrotoluene	tr	C 7 H 6 N 2 O 4	1.26E+01	3.24E+00	2.31E+00	1.47E+00	3.41E+00	9.40E+01	3.40E+01	8.29E+01	9.72E+03	−1.96E+04	5.01E+00
2,4-dinitrotoluene	tr	C 7 H 6 N 2 O 4	1.26E+01	3.23E+00	2.33E+00	1.48E+00	8.29E−01	9.40E+01	3.40E+01	8.33E+01	9.30E+03	−1.88E+04	3.75E+00
2,5-dinitrotoluene	tr	C 7 H 6 N 2 O 4	1.26E+01	3.23E+00	2.33E+00	1.46E+00	8.77E−01	9.40E+01	3.40E+01	8.40E+01	9.24E+03	−1.87E+04	5.15E+00
2,6-dinitrotoluene	tr	C 7 H 6 N 2 O 4	1.26E+01	3.24E+00	2.28E+00	1.53E+00	9.15E−01	9.40E+01	3.40E+01	8.29E+01	9.45E+03	−1.91E+04	3.99E+00
3,4-dinitrotoluene	tr	C 7 H 6 N 2 O 4	1.26E+01	3.23E+00	2.35E+00	1.41E+00	8.45E−01	9.40E+01	3.40E+01	8.37E+01	9.49E+03	−1.92E+04	5.08E+00
3,5-dinitrotoluene	tr	C 7 H 6 N 2 O 4	1.26E+01	3.23E+00	2.38E+00	1.36E+00	8.70E−01	9.40E+01	3.40E+01	8.27E+01	9.19E+03	−1.86E+04	3.91E+00
2-methyl-4,6-dinitrophenol	tr	C 7 H 6 N 2 O 5	1.30E+01	3.37E+00	2.49E+00	1.56E+00	9.01E−01	1.02E+02	3.70E+01	8.86E+01	1.08E+04	−2.19E+04	5.00E+00
3-chlorotoluene	tr	C 7 H 7 Cl 1	1.24E+01	2.89E+00	2.23E+00	1.21E+00	8.50E−01	6.60E+01	2.10E+01	6.38E+01	3.94E+03	−7.93E+03	3.84E+00
4-chlorotoluene	tr	C 7 H 7 Cl 1	1.24E+01	2.89E+00	2.23E+00	1.26E+00	6.59E−01	6.60E+01	2.10E+01	6.42E+01	3.93E+03	−7.91E+03	4.33E+00
2-nitrotoluene	tr	C 7 H 7 N 1 O 2	1.26E+01	2.88E+00	2.01E+00	1.28E+00	7.52E−01	7.20E+01	2.60E+01	6.87E+01	6.22E+03	−1.21E+04	3.57E+00
4-nitrotoluene	tr	C 7 H 7 N 1 O 2	1.25E+01	2.82E+00	2.01E+00	1.19E+00	6.36E−01	7.20E+01	2.60E+01	6.94E+01	5.82E+03	−1.18E+04	3.76E+00
2-methyl-3,5-dinitroaniline	tr	C 7 H 7 N 3 O 4	1.41E+01	3.44E+00	2.54E+00	1.69E+00	9.15E−01	1.02E+02	3.70E+01	9.49E+01	1.08E+04	−2.18E+04	4.12E+00
2-methyl-3,6-dinitroaniline	tr	C 7 H 7 N 3 O 4	1.41E+01	3.44E+00	2.50E+00	1.73E+00	9.71E−01	1.02E+02	3.70E+01	9.73E+01	1.09E+04	−2.20E+04	5.34E+00
5-methyl-2,4-dinitroaniline	tr	C 7 H 7 N 3 O 4	1.41E+01	3.44E+00	2.56E+00	1.61E+00	9.79E−01	1.02E+02	3.70E+01	9.77E+01	1.09E+04	−2.20E+04	4.92E+00
4-methyl-2,6-dinitroaniline	tr	C 7 H 7 N 3 O 4	1.41E+01	3.44E+00	2.58E+00	1.56E+00	9.66E−01	1.02E+02	3.70E+01	9.81E+01	1.09E+04	−2.21E+04	4.21E+00
4-methyl-3,5-dinitroaniline	tr	C 7 H 7 N 3 O 4	1.41E+01	3.44E+00	2.54E+00	1.64E+00	9.85E−01	1.02E+02	3.70E+01	9.57E+01	1.09E+04	−2.20E+04	4.46E+00
4-chloro-3-methyl-phenol	tr	C 7 H 7 O 1 Cl 1	1.28E+01	3.03E+00	2.34E+00	1.56E+00	7.46E−01	7.40E+01	2.40E+01	6.88E+01	5.03E+03	−1.01E+04	4.27E+00
toluene	tr	C 7 H 8	1.24E+01	2.41E+00	1.65E+00	9.40E−01	5.34E−01	5.00E+01	1.80E+01	4.56E+01	3.23E+03	−6.27E+03	3.32E+00
2-methyl-3-nitroaniline	tr	C 7 H 8 N 2 O 2	1.40E+01	3.03E+00	2.17E+00	1.48E+00	7.91E−01	8.00E+01	2.90E+01	7.94E+01	7.23E+03	−1.46E+04	3.48E+00
2-methyl-4-nitroaniline	tr	C 7 H 8 N 2 O 2	1.40E+01	3.02E+00	2.22E+00	1.41E+00	7.35E−01	8.00E+01	2.90E+01	8.18E+01	7.06E+03	−1.43E+04	3.24E+00
2-methyl-6-nitroaniline	tr	C 7 H 8 N 2 O 2	1.40E+01	3.03E+00	2.19E+00	1.44E+00	7.79E−01	8.00E+01	2.90E+01	7.98E+01	7.06E+03	−1.43E+04	3.80E+00
3-methyl-6-nitroaniline	tr	C 7 H 8 N 2 O 2	1.40E+01	3.02E+00	2.24E+00	1.32E+00	8.04E−01	8.00E+01	2.90E+01	8.17E+01	7.16E+03	−1.45E+04	3.80E+00
4-methyl-2-nitroaniline	tr	C 7 H 8 N 2 O 2	1.40E+01	3.02E+00	2.24E+00	1.32E+00	7.66E−01	8.00E+01	2.90E+01	8.19E+01	7.19E+03	−1.45E+04	3.79E+00
4-methyl-3-nitroaniline	tr	C 7 H 8 N 2 O 2	1.40E+01	3.02E+00	2.23E+00	1.36E+00	7.75E−01	8.00E+01	2.90E+01	8.05E+01	7.17E+03	−1.45E+04	3.77E+00
o-cresol (2-hydroxytoluene)	tr	C 7 H 8 O 1	1.28E+01	2.55E+00	1.79E+00	1.12E+00	5.63E−01	5.80E+01	2.10E+01	5.90E+01	4.27E+03	−8.30E+03	3.77E+00

Table 1. (Continued)

compd	ID	formula	information										exp
			$^0\chi$	$^1\chi$	$^2\chi$	$^3\chi$	$^4\chi$	N	NFL	AP	NNR	ENA	$-\log(LC_{50})$
m-cresol (3-hydroxytoluene)	tr	C 7 H 8 O 1	1.28E+01	2.54E+00	1.84E+00	1.00E+00	6.28E−01	5.80E+01	2.10E+01	5.92E+01	4.21E+03	−8.19E+03	3.29E+00
3-hydroxyanisole	tr	C 7 H 8 O 2	1.32E+01	2.66E+00	1.70E+00	1.05E+00	6.10E−01	6.60E+01	2.40E+01	6.48E+01	5.15E+03	−1.04E+04	3.21E+00
1,2-dimethylbenzene	tr	C 8 H 10	1.53E+01	2.83E+00	2.08E+00	1.43E+00	6.63E−01	5.80E+01	2.10E+01	6.25E+01	4.10E+03	−8.25E+03	3.48E+00
1,4-dimethylbenzene	tr	C 8 H 10	1.53E+01	2.82E+00	2.15E+00	1.22E+00	6.37E−01	5.80E+01	2.10E+01	6.30E+01	4.04E+03	−8.12E+03	4.21E+00
2,4-dimethyl-phenol	tr	C 8 H 10 O 1	1.57E+01	2.96E+00	2.35E+00	1.21E+00	9.51E−01	6.60E+01	2.40E+01	6.78E+01	5.14E+03	−1.03E+04	3.86E+00
3,4-dimethyl-phenol	tr	C 8 H 10 O 1	1.57E+01	2.96E+00	2.27E+00	1.49E+00	7.26E−01	6.60E+01	2.40E+01	6.80E+01	5.14E+03	−1.03E+04	3.90E+00
1,4-dimethoxybenzene	tr	C 8 H 10 O 2	1.61E+01	3.05E+00	1.88E+00	1.30E+00	7.12E−01	7.40E+01	2.70E+01	7.46E+01	6.23E+03	−1.25E+04	3.07E+00
1,2,4-trimethyl-benzene	tr	C 9 H 12	1.82E+01	3.24E+00	2.59E+00	1.66E+00	8.91E−01	6.60E+01	2.40E+01	7.12E+01	5.26E+03	−1.03E+04	4.21E+00
1,2,4,5-tetrachlorobenzene	te	C 6 H 2 Cl 4	9.68E+00	3.92E+00	3.31E+00	2.40E+00	1.26E+00	1.06E+02	2.70E+01	8.73E+01	5.89E+03	−1.18E+04	5.85E+00
1,3-dichlorobenzene	te	C 6 H 4 Cl 2	9.57E+00	2.95E+00	2.31E+00	1.26E+00	8.95E−01	7.40E+01	2.10E+01	6.55E+01	3.95E+03	−7.72E+03	4.30E+00
2,4-dichlorotoluene	te	C 7 H 6 Cl 2	1.25E+01	3.37E+00	2.74E+00	1.77E+00	9.75E−01	8.20E+01	2.40E+01	7.40E+01	4.91E+03	−9.87E+03	4.54E+00
3-nitrotoluene	te	C 7 H 7 N 1 O 2	1.26E+01	2.87E+00	2.07E+00	1.19E+00	7.24E−01	7.20E+01	2.60E+01	6.86E+01	5.85E+03	−1.18E+04	3.63E+00
3-methyl-2,4-dinitroaniline	te	C 7 H 7 N 3 O 4	1.41E+01	3.44E+00	2.51E+00	1.67E+00	1.05E+00	1.02E+02	3.70E+01	9.71E+01	1.10E+04	−2.23E+04	4.26E+00
5-methyl-2,6-dinitroaniline	te	C 7 H 7 N 3 O 4	1.41E+01	3.44E+00	2.53E+00	1.64E+00	1.02E+00	1.02E+02	3.70E+01	9.80E+01	1.10E+04	−2.23E+04	4.18E+00
2-methyl-5-nitroaniline	te	C 7 H 8 N 2 O 2	1.40E+01	3.02E+00	2.22E+00	1.42E+00	7.07E−01	8.00E+01	2.90E+01	8.08E+01	7.25E+03	−1.46E+04	3.35E+00
p-cresol (4-hydroxytoluene)	te	C 7 H 8 O 1	1.28E+01	2.54E+00	1.84E+00	1.03E+00	5.45E−01	5.80E+01	2.10E+01	5.96E+01	4.05E+03	−8.16E+03	3.58E+00
2,6-dimethyl-phenol	te	C 8 H 10 O 1	1.57E+01	2.97E+00	2.24E+00	1.44E+00	8.06E−01	6.60E+01	2.40E+01	6.74E+01	5.22E+03	−1.05E+04	3.75E+00
2,4,6-trichlorophenol	te	C 6 H 3 O 1 Cl 3	1.00E+01	3.58E+00	2.97E+00	1.77E+00	1.35E+00	9.80E+01	2.70E+01	8.12E+01	6.02E+03	−1.21E+04	4.33E+00

[a] $^{0-4}\chi$ = valence connectivity indices; N = sum of atomic numbers; NFL = number of filled levels; AP = average polarizability (PM3); NNR = nuclear−nuclear repulsion; ENA = electron−nuclear attraction.

descriptors. This network was proposed by Kohonen[7,8] in 1982, as an algorithm able to classify data (*N*-dimensional vectors, with *N* usually very large) via a projection of these data into a subspace of lower dimension *M* (usually $M = 2$), called map, preserving its topology in the original space. In this context, the word topology is used instead of geometry to denote relative distance between points in a certain space. The dynamical process that occurs to structure the topology of data in the maps is known as self-organization and is inspired by the organization of cognitive functions in different lobes of the brain.[10]

The algorithm is an optimization process in which the weights associated to each node or neuron in a two-dimensional lattice are adjusted to cluster the input information while preserving the topology of the original data. Thus, the weights are vectors with the same dimension as the input data, which are initiated with random values. The map used in the current study to cluster the two sets of data according to toxicity and molecular features of each chemical was a two-dimensional 7×7 grid with hexagonal lattices. This dimension was selected after analyzing the quality of the clustering and the number of empty nodes for map sizes ranging from 5×5 to 10×10. The resulting weights define the cluster or vector centers that sample the input space when a sufficient number of input vectors selected randomly are sequentially presented to the network. The weights are associated with the input variables so that close nodes are sensitive to inputs that have similar representations of their characteristics.

The standard algorithm proposed by Kohonen operates in the following sequence:

1. Presentation of an input vector x_i of dimension *N* to the network;

2. Calculation of the Euclidian distance between this input vector and all nodes in the network lattice

$$d_j = \sum_{i=0}^{N-1} (x_i(t) - w_{ij}(t))^2 \qquad (1)$$

In this equation $x_i(t)$ is the *i*th component of the *N*-dimensional input vector and $w_{ij}(t)$ the connection strength (weight) between the input neuron *i* and the mapping array node *j* at time (position) *t* in the sequence of total data presentation to the network. Each of these presentations are known as epoch;

3. Selection of the node with mininum distance, *j**. This node is the winner neuron or best matching unit (BMU);

4. Update weights of node *j** and neighbors, restricted to the neighborhood $N_{j^*}(t)$

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)) \qquad (2)$$

for $j \in N_{j^*}(t)$ and $1 \le i \le N$. Here $\eta(t)$ is a function that decreases monotonically over the environment of the winner neuron. It defines the region of influence that the input vector has on SOM. The function $\eta(t)$ is defined by the neighborhood function η_o and the learning rate $\alpha(t)$ according to

$$\eta(t) = \eta_o(||r_c - r||,t)\alpha(t) \qquad (3)$$

where *r* is the location of the units or neurons on the grid of the map.

Table 2. Best Set of Molecular Descriptors and Experimental Oral Rat Toxicity LD_{50} for a Heterogeneous Set of 155 Organic Compounds[a]

	information												
compd	ID	formula	$^0\chi$	$^1\chi$	$^4\chi$	EE	HP	NNR	ENA	RE	NFL	κ	exp log(LD_{50})
tetrachloromethane	tr	C 1 Cl 4	5.03E+00	2.26E+00	0.00E+00	−1.84E+01	0.00E+00	2.22E+03	−4.37E+03	−4.08E+01	1.60E+01	1.00E+00	3.37E+00
tribromomethane	tr	C 1 H 1 Br 3	7.46E+00	3.40E+00	0.00E+00	−1.83E+01	4.11E+00	1.43E+03	−2.81E+03	−3.48E+01	1.30E+01	1.33E+00	3.06E+00
trichloromethane	tr	C 1 H 1 Cl 3	4.97E+00	1.96E+00	0.00E+00	−1.94E+01	3.10E+00	1.47E+03	−2.87E+03	−4.07E+01	1.30E+01	1.33E+00	2.96E+00
triiodomethane	tr	C 1 H 1 I 3	9.08E+00	4.33E+00	0.00E+00	−1.77E+01	5.84E+00	1.36E+03	−2.68E+03	−3.15E+01	1.30E+01	1.33E+00	2.55E+00
dichloromethane	tr	C 1 H 2 Cl 2	4.97E+00	1.60E+00	0.00E+00	−2.05E+01	6.30E+00	8.81E+02	−1.71E+03	−4.09E+01	1.00E+01	2.00E+00	3.20E+00
formaldehyde	tr	C 1 H 2 O 1	3.12E+00	2.89E−01	0.00E+00	−2.17E+01	1.88E+01	3.92E+02	−7.24E+02	−4.91E+01	6.00E+00	0.00E+00	2.90E+00
formic acid	tr	C 1 H 2 O 2	3.43E+00	4.94E−01	0.00E+00	−2.74E+01	1.21E+01	9.05E+02	−1.70E+03	−7.09E+01	9.00E+00	2.00E+00	3.04E+00
bromomethane	tr	C 1 H 3 Br 1	5.96E+00	1.96E+00	0.00E+00	−2.12E+01	4.02E+00	4.38E+02	−8.35E+02	−3.87E+01	7.00E+00	0.00E+00	2.33E+00
chloromethane	tr	C 1 H 3 Cl 1	5.13E+00	1.13E+00	0.00E+00	−2.17E+01	6.14E+00	4.59E+02	−8.76E+02	−4.12E+01	7.00E+00	0.00E+00	3.26E+00
nitromethane	tr	C 1 H 3 N 1 O 2	5.26E+00	7.57E−01	0.00E+00	−3.86E+01	1.82E+01	1.61E+03	−3.06E+03	−9.50E+01	1.20E+01	1.33E+00	2.97E+00
methanol	tr	C 1 H 4 O 1	5.45E+00	4.47E−01	0.00E+00	−2.82E+01	1.23E+01	5.73E+02	−1.07E+03	−6.14E+01	7.00E+00	0.00E+00	3.75E+00
nitrotrichloromethane	tr	C 1 N 1 O 2 Cl 3	5.16E+00	2.25E+00	0.00E+00	−3.52E+01	1.25E+01	3.94E+03	−7.99E+03	−9.37E+01	2.10E+01	1.85E+00	2.40E+00
1,2,4,5-tetramethylbenzene	tr	C 10 H 14	2.12E+01	3.65E+00	1.10E+00	−7.80E+01	1.01E+00	6.27E+03	−1.23E+04	−1.95E+02	2.10E+01	2.94E+00	3.84E+00
sec-butyl-benzene	tr	C 10 H 14	2.07E+01	3.89E+00	1.02E+00	−1.47E+02	5.13E−01	6.41E+03	−1.25E+04	−2.97E+02	2.70E+01	4.00E+00	3.35E+00
1-decanol	tr	C 10 H 22 O 1	2.98E+01	5.02E+00	1.25E+00	−1.78E+02	2.69E+00	8.33E+03	−1.63E+04	−3.56E+02	3.40E+01	1.00E+01	3.67E+00
1-methyl naphthalene	tr	C 11 H 10	1.65E+01	3.82E+00	1.39E+00	−1.45E+02	1.03E+00	6.52E+03	−1.26E+03	−3.01E+02	2.70E+01	3.16E+00	3.26E+00
1-dodecanol	tr	C 12 H 26 O 1	3.52E+01	6.02E+00	1.60E+00	−2.11E+02	3.32E+00	1.03E+04	−2.08E+04	−4.21E+02	4.00E+01	1.20E+01	4.11E+00
1-tetradecanol	tr	C 14 H 30 O 1	4.06E+01	7.02E+00	1.95E+00	−2.45E+02	2.90E+00	1.27E+04	−2.55E+04	−4.87E+02	4.60E+01	1.40E+01	4.52E+00
1-hexadecanol	tr	C 16 H 34 O 1	4.61E+01	8.02E+00	2.31E+00	−2.78E+02	2.58E+00	1.52E+04	−3.04E+04	−5.52E+02	5.20E+01	1.60E+01	3.70E+00
1-heptadecanol	tr	C 17 H 36 O 1	4.88E+01	8.52E+00	2.48E+00	−2.95E+02	2.44E+00	1.64E+04	−3.29E+04	−5.85E+02	5.50E+01	1.70E+01	4.71E+00
tetrachloroethene	tr	C 2 Cl 4	5.53E+00	2.51E+00	0.00E+00	−7.96E+01	6.43E+00	2.77E+03	−5.45E+03	−6.68E+01	1.80E+01	2.22E+00	3.42E+00
1,1,2-trichlorotrifluoroethane	tr	C 2 F 3 Cl 3	5.53E+00	2.52E+00	0.00E+00	−3.36E+01	1.61E+00	5.94E+03	−1.09E+04	−9.29E+01	2.50E+01	1.75E+00	4.63E+00
trichloroethene	tr	C 2 H 1 Cl 3	5.47E+00	2.07E+00	0.00E+00	−3.03E+01	1.36E+00	1.96E+03	−3.85E+03	−6.55E+01	1.50E+01	2.25E+00	3.69E+00
pentachloroethane	tr	C 2 H 1 Cl 5	7.74E+00	3.29E+00	0.00E+00	−3.37E+01	4.85E+00	4.10E+03	−8.24E+03	−7.29E+01	2.20E+01	1.85E+00	2.96E+00
1,1,2,2-tetrabromoethane	tr	C 2 H 2 Br 4	1.10E+01	4.86E+00	0.00E+00	−3.31E+01	5.14E+00	3.05E+03	−6.02E+03	−6.50E+01	1.90E+01	2.22E+00	3.08E+00
1,1,2,2-tetrachloroethane	tr	C 2 H 2 Cl 4	7.68E+00	2.95E+00	0.00E+00	−3.32E+01	5.19E+00	5.47E+03	−1.08E+04	−8.54E+01	2.50E+01	2.22E+00	2.30E+00
chloroethene	tr	C 2 H 3 Cl 1	5.42E+00	1.06E+00	0.00E+00	−3.20E+01	1.66E+00	7.90E+02	−1.51E+03	−6.34E+01	9.00E+00	2.00E+00	2.70E+00
1,1,1-trichloroethane	tr	C 2 H 3 Cl 3	7.90E+00	2.20E+00	0.00E+00	−3.59E+01	4.31E+00	2.27E+03	−4.45E+03	−7.31E+01	1.60E+01	1.00E+00	3.98E+00
1,1,2-trichloroethane	tr	C 2 H 3 Cl 3	7.68E+00	2.52E+00	0.00E+00	−3.61E+01	6.32E+00	2.34E+03	−4.58E+03	−7.38E+01	1.60E+01	2.25E+00	2.92E+00
acetonitrile	tr	C 2 H 3 N 1	4.95E+00	7.24E−01	0.00E+00	−3.71E+01	1.80E+01	6.89E+02	−1.29E+03	−7.74E+01	8.00E+00	2.00E+00	3.44E+00
2,2,2-trifluoroethanol	tr	C 2 H 3 O 1 F 3	5.79E+00	1.24E+00	0.00E+00	−4.25E+01	6.44E+01	3.54E+03	−6.90E+03	−1.15E+02	1.90E+01	1.63E+00	2.38E+00
1,2-dibromoethane	tr	C 2 H 4 Br 2	9.34E+00	3.27E+00	0.00E+00	−3.60E+01	6.75E+00	1.43E+03	−2.89E+03	−6.90E+01	1.30E+01	3.00E+00	2.03E+00
1-bromo-2-chloroethane	tr	C 2 H 4 Cl 1 Br 1	8.51E+00	2.69E+00	0.00E+00	−3.65E+01	7.23E+00	1.51E+03	−2.94E+03	−7.12E+01	1.30E+01	3.00E+00	1.81E+00
1,2-dichloroethane	tr	C 2 H 4 Cl 2	7.68E+00	2.10E+00	0.00E+00	−3.70E+01	7.75E+00	1.54E+03	−2.99E+03	−7.34E+01	1.30E+01	3.00E+00	2.83E+00
acetaldehyde	tr	C 2 H 4 O 1	5.99E+00	8.13E−01	0.00E+00	−3.84E+01	8.00E+00	8.93E+02	−1.69E+03	−8.20E+01	9.00E+00	2.00E+00	2.82E+00
ethylene oxide (oxirane)	tr	C 2 H 4 O 1	5.82E+00	1.08E+00	2.20E−01	−3.80E+01	1.34E+01	8.89E+02	−1.81E+03	−8.12E+01	9.00E+00	2.22E−01	1.86E+00
acetic acid	tr	C 2 H 4 O 2	6.36E+00	9.28E−01	0.00E+00	−4.39E+01	7.91E+00	1.57E+03	−3.01E+03	−1.04E+02	1.20E+01	1.33E+00	3.52E+00
methyl formate	tr	C 2 H 4 O 2	6.39E+00	8.80E−01	0.00E+00	−4.40E+01	9.21E+00	1.57E+03	−3.00E+03	−1.02E+02	1.20E+01	3.00E+00	3.20E+00
bromoethane	tr	C 2 H 5 Br 1	8.67E+00	2.09E+00	0.00E+00	−3.78E+01	3.12E+00	9.87E+02	−1.90E+03	−7.18E+01	1.00E+01	2.00E+00	3.13E+00
chloroethane	tr	C 2 H 5 Cl 1	7.84E+00	1.51E+00	0.00E+00	−3.83E+01	9.39E+00	1.01E+03	−1.94E+03	−7.40E+01	1.00E+01	2.00E+00	3.26E+00
nitroethane	tr	C 2 H 5 N 1 O 2	7.97E+00	1.35E+00	0.00E+00	−5.52E+01	1.52E+01	2.25E+03	−4.59E+03	−1.28E+02	1.50E+01	2.25E+00	3.04E+00
ethanol	tr	C 2 H 6 O 1	8.15E+00	1.02E+00	0.00E+00	−4.49E+01	8.80E+00	1.13E+03	−2.16E+03	−9.39E+01	1.00E+01	2.00E+00	3.85E+00
dimethyl sulfoxide	tr	C 2 H 6 O 1 S 1	9.28E+00	2.09E+00	0.00E+00	−4.85E+01	1.64E+01	1.57E+03	−3.19E+03	−9.76E+01	1.30E+01	1.33E+00	4.16E+00
dimethyl sulfide	tr	C 2 H 6 S 1	9.22E+00	2.44E+00	0.00E+00	−4.31E+01	4.54E+00	9.90E+02	−1.89E+03	−8.25E+01	1.00E+01	2.00E+00	3.52E+00
ethanethiol	tr	C 2 H 6 S 1	9.05E+00	1.65E+00	0.00E+00	−3.81E+01	6.59E+00	8.51E+02	−1.73E+03	−7.47E+01	9.00E+00	2.00E+00	3.29E+00
1,3-dichloropropene	tr	C 3 H 4 Cl 2	8.13E+00	2.20E+00	3.02E−01	−4.74E+01	7.58E+00	1.90E+03	−3.84E+03	−9.61E+01	1.50E+01	4.00E+00	2.67E+00
3-chloro-1-propene	tr	C 3 H 5 Cl 1	8.12E+00	1.62E+00	0.00E+00	−4.84E+01	8.63E+00	1.40E+03	−2.70E+03	−9.53E+01	1.20E+01	3.00E+00	2.85E+00
1,2,3-trichloropropane	tr	C 3 H 5 Cl 3	1.04E+01	3.07E+00	3.70E−01	−5.25E+01	7.43E+00	3.17E+03	−6.21E+03	−1.06E+02	1.20E+01	3.20E+00	2.51E+00
propanenitrile	tr	C 3 H 5 N 1	7.65E+00	1.28E+00	0.00E+00	−5.37E+01	1.43E+01	1.20E+03	−2.42E+03	−1.10E+02	1.10E+01	3.00E+00	1.59E+00
1,2-dibromopropane	tr	C 3 H 6 Br 2	1.22E+01	3.50E+00	0.00E+00	−5.27E+01	5.78E+00	2.25E+03	−4.54E+03	−1.02E+02	1.60E+01	2.25E+00	2.87E+00
1,3-dibromopropane	tr	C 3 H 6 Br 2	1.20E+01	3.77E+00	1.36E+00	−5.27E+01	3.36E+00	2.16E+03	−4.36E+03	−1.02E+02	1.60E+01	4.00E+00	2.87E+00
1,2-dichloropropane	tr	C 3 H 6 Cl 2	1.05E+01	2.44E+00	0.00E+00	−5.37E+01	6.55E+00	2.39E+03	−4.66E+03	−1.06E+02	1.60E+01	2.25E+00	3.29E+00

Table 2. (Continued)

	information												
compd	ID	formula	$^0\chi$	$^1\chi$	$^4\chi$	EE	HP	NNR	ENA	RE	NFL	κ	exp log(LD_{50})
acetone	tr	C 3 H 6 O 1	8.91E+00	1.20E+00	0.00E+00	−5.51E+01	1.04E+01	1.47E+03	−2.98E+03	−1.15E+02	1.20E+01	1.33E+00	3.76E+00
propionaldehyde	tr	C 3 H 6 O 1	8.69E+00	1.35E+00	0.00E+00	−5.50E+01	6.34E+00	1.52E+03	−2.91E+03	−1.15E+02	1.20E+01	3.00E+00	3.15E+00
ethyl formate	tr	C 3 H 6 O 2	9.10E+00	1.47E+00	6.80E−02	−6.06E+01	7.23E+00	2.31E+03	−4.46E+03	−1.35E+02	1.50E+01	4.00E+00	3.27E+00
N,*N*-dimethylformamide	tr	C 3 H 7 N 1 O	1.04E+01	1.39E+00	0.00E+00	−6.60E+01	1.37E+01	2.35E+03	−4.53E+03	−1.43E+02	1.50E+01	2.25E+00	3.45E+00
1-nitropropane	tr	C 3 H 7 N 1 O 2	1.07E+01	1.85E+00	1.75E−01	−7.19E+01	1.20E+01	3.23E+03	−6.22E+03	−1.60E+02	1.80E+01	3.20E+00	2.66E+00
1-propanol	tr	C 3 H 8 O 1	1.09E+01	1.52E+00	0.00E+00	−6.15E+01	6.87E+00	1.80E+03	−3.46E+03	−1.27E+02	1.30E+01	3.00E+00	3.27E+00
1-butanol	tr	C 4 H 10 O 1	1.36E+01	2.02E+00	1.58E−01	−7.82E+01	5.65E+00	2.45E+03	−4.94E+03	−1.59E+02	1.60E+01	4.00E+00	2.90E+00
2-butanol	tr	C 4 H 10 O 1	1.37E+01	1.95E+00	0.00E+00	−7.81E+01	5.71E+00	2.65E+03	−5.14E+03	−1.59E+02	1.60E+01	2.25E+00	3.81E+00
2-methyl-1-propanol	tr	C 4 H 10 O 1	1.37E+01	1.88E+00	0.00E+00	−7.82E+01	5.74E+00	2.55E+03	−5.14E+03	−1.59E+02	1.60E+01	2.25E+00	3.39E+00
diethyl ether	tr	C 4 H 10 O 1	1.38E+01	1.99E+00	2.04E−01	−7.82E+01	2.94E+00	2.60E+03	−5.05E+03	−1.58E+02	1.60E+01	4.00E+00	3.08E+00
tert-butyl alcohol	tr	C 4 H 10 O 1	1.39E+01	1.72E+00	0.00E+00	−7.82E+01	5.60E+00	2.72E+03	−5.28E+03	−1.59E+02	1.60E+01	1.00E+00	3.54E+00
1-butanethiol	tr	C 4 H 10 S 1	1.45E+01	2.65E+00	4.73E−01	−7.14E+01	4.65E+00	2.27E+03	−4.40E+03	−1.40E+02	1.50E+01	4.00E+00	3.18E+00
diethyl sulfide	tr	C 4 H 10 S 1	1.46E+01	3.14E+00	6.11E−01	−7.64E+01	3.10E+00	2.35E+03	−4.56E+03	−1.48E+02	1.60E+01	4.00E+00	3.77E+00
diethylamine	tr	C 4 H 11 N 1	1.49E+01	2.12E+00	2.50E−01	−8.35E+01	2.28E+00	2.56E+03	−4.96E+03	−1.67E+02	1.60E+01	4.00E+00	2.73E+00
thiophene	tr	C 4 H 4 S 1	7.53E+00	2.41E+00	6.79E−01	−5.78E+01	0.00E+00	1.72E+03	−3.32E+03	−1.18E+02	1.30E+01	1.44E+00	3.15E+00
butyraldehyde	tr	C 4 H 8 O 1	1.14E+01	1.85E+00	1.18E−01	−7.17E+01	5.28E+00	2.24E+03	−4.32E+03	−1.47E+02	1.50E+01	4.00E+00	3.40E+00
tetrahydrofuran	tr	C 4 H 8 O 1	1.12E+01	2.08E+00	5.10E−01	−7.18E+01	5.69E+00	2.39E+03	−4.81E+03	−1.49E+02	1.50E+01	1.44E+00	3.22E+00
ethyl acetate	tr	C 4 H 8 O 2	1.20E+01	1.90E+00	2.03E−01	−7.72E+01	5.35E+00	3.22E+03	−6.26E+03	−1.68E+02	1.80E+01	3.20E+00	3.75E+00
1-chlorobutane	tr	C 4 H 9 Cl 1	1.33E+01	2.51E+00	4.00E−01	−7.17E+01	5.54E+00	2.42E+03	−4.70E+03	−1.39E+02	1.60E+01	4.00E+00	3.43E+00
2-chlorobutane	tr	C 4 H 9 Cl 1	1.34E+01	2.35E+00	0.00E+00	−7.17E+01	5.50E+00	2.51E+03	−4.88E+03	−1.39E+02	1.60E+01	2.25E+00	4.24E+00
N,*N*-dimethyl acetamide	tr	C 4 H 9 N 1 O 1	1.34E+01	1.82E+00	0.00E+00	−8.27E+01	1.15E+01	3.15E+03	−6.35E+03	−1.76E+02	1.80E+01	2.22E+00	3.69E+00
2-pentanone	tr	C 5 H 10 O 1	1.43E+01	2.26E+00	3.52E−01	−8.84E+01	7.73E+00	3.13E+03	−6.08E+03	−1.81E+02	1.80E+01	3.20E+00	3.20E+00
3-methyl-2-butanone	tr	C 5 H 10 O 1	1.45E+01	2.15E+00	0.00E+00	−8.85E+01	7.26E+00	3.24E+03	−6.30E+03	−1.80E+02	1.80E+01	2.22E+00	2.17E+00
pentanal	tr	C 5 H 10 O 1	1.41E+01	2.35E+00	2.87E−01	−8.83E+01	4.46E+00	2.92E+03	−5.90E+03	−1.80E+02	1.80E+01	5.00E+00	3.51E+00
tetrahydropyran	tr	C 5 H 10 O 1	1.39E+01	2.58E+00	6.87E−01	−8.85E+01	4.66E+00	3.32E+03	−6.68E+03	−1.82E+02	1.80E+01	2.22E+00	3.48E+00
isopropyl acetate	tr	C 5 H 10 O 2	1.49E+01	2.30E+00	3.32E−01	−9.38E+01	4.15E+00	4.13E+03	−8.32E+03	−2.00E+02	2.10E+01	3.06E+00	3.48E+00
1-pentanol	tr	C 5 H 12 O 1	1.63E+01	2.52E+00	3.62E−01	−9.49E+01	4.50E+00	3.38E+03	−6.57E+03	−1.92E+02	1.90E+01	5.00E+00	3.48E+00
3-methyl-1-butanol	tr	C 5 H 12 O 1	1.64E+01	2.38E+00	2.58E−01	−9.48E+01	4.55E+00	3.39E+03	−6.82E+03	−1.92E+02	1.90E+01	3.20E+00	3.11E+00
pyridine	tr	C 5 H 5 N 1	8.33E+00	1.85E+00	3.13E−01	−7.41E+01	8.84E+00	2.38E+03	−4.60E+03	−1.58E+02	1.50E+01	2.22E+00	2.95E+00
cyclopentene	tr	C 5 H 8	1.13E+01	2.15E+00	5.89E−01	−7.68E+01	1.28E+00	2.14E+03	−4.13E+03	−1.52E+02	1.40E+01	1.44E+00	3.22E+00
N-methyl-2-pyrrolidinone	tr	C 5 H 9 N 1 O 1	1.35E+01	2.54E+00	7.83E−01	−9.29E+01	1.23E+01	3.87E+03	−7.80E+03	−1.99E+02	2.00E+01	1.85E+00	3.59E+00
hexachlorobenzene	tr	C 6 Cl 6	9.79E+00	4.90E+00	2.00E+00	−7.58E+01	8.60E+00	8.58E+03	−1.70E+04	−1.72E+02	3.30E+01	3.40E+00	4.00E+00
cyclohexanone	tr	C 6 H 10 O 1	1.44E+01	2.91E+00	9.46E−01	−9.88E+01	6.30E+00	4.00E+03	−7.80E+03	−2.04E+02	2.00E+01	2.34E+00	3.19E+00
cyclohexane	tr	C 6 H 12	1.62E+01	3.00E+00	1.06E+00	−1.00E+02	0.00E+00	3.38E+03	−6.58E+03	−1.97E+02	1.80E+01	2.22E+00	4.47E+00
1-hexen-3-ol	tr	C 6 H 12 O 1	1.67E+01	2.62E+00	4.14E−01	−1.05E+02	4.65E+00	3.97E+03	−8.00E+03	−2.13E+02	2.10E+01	4.17E+00	3.67E+00
2-hexanone	tr	C 6 H 12 O 1	1.70E+01	2.76E+00	4.26E−01	−1.05E+02	6.77E+00	4.02E+03	−7.83E+03	−2.13E+02	2.10E+01	4.17E+00	3.41E+00
3,3-dimethyl-2-butanone	tr	C 6 H 12 O 1	1.74E+01	2.45E+00	0.00E+00	−1.05E+02	6.15E+00	4.34E+03	−8.48E+03	−2.13E+02	2.10E+01	1.85E+00	2.79E+00
4-methyl-2-pentanone	tr	C 6 H 12 O 1	1.72E+01	2.62E+00	5.75E−01	−1.05E+02	6.78E+00	4.17E+03	−8.12E+03	−2.13E+02	2.10E+01	3.06E+00	3.32E+00
hexanal	tr	C 6 H 12 O 1	1.68E+01	2.85E+00	4.78E−01	−1.05E+02	3.86E+00	3.91E+03	−7.61E+03	−2.13E+02	2.10E+01	6.00E+00	3.69E+00
ethyl butyrate	tr	C 6 H 12 O 2	1.74E+01	2.96E+00	4.37E−01	−1.10E+02	4.13E+00	5.13E+03	−1.00E+04	−2.33E+02	2.40E+01	5.14E+00	4.11E+00
hexanoic acid	tr	C 6 H 12 O 2	1.72E+01	2.99E+00	5.26E−01	−1.11E+02	0.00E+00	5.00E+03	−9.73E+03	−2.35E+02	2.40E+01	5.14E+00	3.48E+00
n-butyl acetate	tr	C 6 H 12 O 2	1.74E+01	2.90E+00	3.60E−01	−1.11E+02	3.64E+00	5.07E+03	−9.89E+03	−2.33E+02	2.40E+01	5.14E+00	4.03E+00
hexane	tr	C 6 H 14	1.88E+01	2.91E+00	5.00E−01	−1.06E+02	0.00E+00	3.37E+03	−6.56E+03	−2.06E+02	1.90E+01	5.00E+00	4.46E+00
1-hexanol	tr	C 6 H 14 O 1	1.90E+01	3.02E+00	5.39E−01	−1.12E+02	3.91E+00	4.14E+03	−8.32E+03	−2.25E+02	2.20E+01	6.00E+00	2.86E+00
3-methyl-3-pentanol	tr	C 6 H 14 O 1	1.94E+01	2.84E+00	2.50E−01	−1.11E+02	5.86E+00	4.72E+03	−9.22E+03	−2.25E+02	2.20E+01	2.34E+00	2.85E+00
4-methyl-2-pentanol	tr	C 6 H 14 O 1	1.93E+01	2.81E+00	6.82E−01	−1.11E+02	6.16E+00	4.44E+03	−8.94E+03	−2.25E+02	2.20E+01	3.06E+00	3.41E+00
diisopropyl ether	tr	C 6 H 14 O 1	1.96E+01	2.78E+00	5.44E−01	−1.11E+02	2.29E+00	4.65E+03	−9.09E+03	−2.23E+02	2.20E+01	3.06E+00	3.93E+00
dipropylamine	tr	C 6 H 15 N 1	2.03E+01	3.12E+00	4.79E−01	−1.17E+02	1.42E+00	4.28E+03	−8.35E+03	−2.32E+02	2.20E+01	6.00E+00	2.97E+00
triethylamine	tr	C 6 H 15 N 1	2.06E+01	3.07E+00	6.71E−01	−1.17E+02	5.59E+00	4.47E+03	−8.98E+03	−2.31E+02	2.20E+01	4.17E+00	2.66E+00
1,2-dichlorobenzene	tr	C 6 H 4 Cl 2	9.57E+00	2.96E+00	7.10E−01	−7.88E+01	6.35E+00	4.01E+03	−7.85E+03	−1.67E+02	2.10E+01	2.52E+00	2.70E+00
bromobenzene	tr	C 6 H 5 Br 1	1.03E+01	2.89E+00	7.19E−01	−7.90E+01	5.51E+00	2.97E+03	−5.98E+03	−1.63E+02	1.80E+01	2.34E+00	3.43E+00
chlorobenzene	tr	C 6 H 5 Cl 1	9.52E+00	2.48E+00	5.60E−01	−7.97E+01	4.27E+00	3.11E+03	−6.06E+03	−1.66E+02	1.80E+01	2.34E+00	3.36E+00
fluorobenzene	tr	C 6 H 5 F 1	8.76E+00	2.10E+00	4.15E−01	−7.98E+01	9.02E+00	3.05E+03	−6.15E+03	−1.72E+02	1.80E+01	2.34E+00	3.64E+00

Table 2 (Continued)

			information										
compd	ID	formula	$^0\chi$	$^1\chi$	$^4\chi$	EE	HP	NNR	ENA	RE	NFL	κ	exp log(LD_{50})
iodobenzene	tr	C 6 H 5 I 1	1.09E+01	3.16E+00	8.23E−01	−7.88E+01	6.39E+00	3.04E+03	−5.92E+03	−1.62E+02	1.80E+01	2.34E+00	3.26E+00
nitrobenzene	tr	C 6 H 5 N 1 O 2	9.65E+00	2.46E+00	5.37E−01	−9.61E+01	9.09E+00	4.98E+03	−9.67E+03	−2.18E+02	2.30E+01	3.24E+00	2.81E+00
benzene	tr	C 6 H 6	9.46E+00	2.00E+00	3.85E−01	−8.06E+01	0.00E+00	2.37E+03	−4.58E+03	−1.65E+02	1.50E+01	2.22E+00	3.52E+00
5-methylfurfural (5-methylfuraldehyde)	tr	C 6 H 6 O 2	1.05E+01	2.34E+00	5.17E−01	−9.16E+01	1.23E+01	4.03E+03	−8.13E+03	−2.03E+02	2.10E+01	2.52E+00	3.34E+00
4-methylpyridine	tr	C 6 H 7 N 1	1.13E+01	2.26E+00	4.26E−01	−9.09E+01	7.32E+00	3.24E+03	−6.29E+03	−1.91E+02	1.80E+01	2.34E+00	3.11E+00
aniline	tr	C 6 H 7 N 1	1.10E+01	2.20E+00	4.53E−01	−9.16E+01	5.16E+00	3.24E+03	−6.29E+03	−1.94E+02	1.80E+01	2.34E+00	2.40E+00
2-heptanone	tr	C 7 H 14 O 1	1.97E+01	3.26E+00	6.24E−01	−1.22E+02	5.97E+00	4.96E+03	−9.67E+03	−2.46E+02	2.40E+01	5.14E+00	3.22E+00
4-heptanone	tr	C 7 H 14 O 1	1.97E+01	3.33E+00	6.83E−01	−1.22E+02	5.97E+00	5.02E+03	−9.80E+03	−2.46E+02	2.40E+01	5.14E+00	3.57E+00
5-methyl-2-hexanone	tr	C 7 H 14 O 1	1.99E+01	3.12E+00	4.92E−01	−1.22E+02	6.03E+00	5.12E+03	−9.99E+03	−2.46E+02	2.40E+01	3.94E+00	3.51E+00
heptanal	tr	C 7 H 14 O 1	1.95E+01	3.35E+00	6.54E−01	−1.22E+02	3.41E+00	4.83E+03	−9.42E+03	−2.46E+02	2.40E+01	7.00E+00	4.15E+00
1-heptanol	tr	C 7 H 16 O 1	2.17E+01	3.52E+00	7.15E−01	−1.28E+02	3.60E+00	5.22E+03	−1.02E+04	−2.58E+02	2.50E+01	7.00E+00	2.70E+00
benzoic acid	tr	C 7 H 6 O 2	1.07E+01	2.59E+00	5.81E−01	−1.02E+02	7.05E+00	4.90E+03	−9.54E+03	−2.27E+02	2.30E+01	3.24E+00	3.23E+00
benzyl chloride	tr	C 7 H 7 Cl 1	1.22E+01	3.06E+00	7.45E−01	−9.60E+01	7.16E+00	3.93E+03	−7.90E+03	−1.98E+02	2.10E+01	3.11E+00	3.09E+00
3-nitrotoluene	tr	C 7 H 7 N 1 O 2	1.26E+01	2.87E+00	7.24E−01	−1.13E+02	7.89E+00	5.85E+03	−1.18E+04	−2.52E+02	2.60E+01	3.41E+00	3.03E+00
2-methoxynitrobenzene (2-nitroanisole)	tr	C 7 H 7 N 1 O 3	1.30E+01	2.99E+00	7.63E−01	−1.18E+02	9.88E+00	7.28E+03	−1.47E+04	−2.72E+02	2.90E+01	4.13E+00	2.87E+00
1−6-heptadiyne	tr	C 7 H 8	1.23E+01	2.28E+00	2.93E−01	−1.04E+02	1.37E+00	3.62E+03	−7.05E+03	−2.08E+02	1.90E+01	6.00E+00	3.36E+00
toluene	tr	C 7 H 8	1.24E+01	2.41E+00	5.34E−01	−9.74E+01	7.49E−01	3.23E+03	−6.27E+03	−1.99E+02	1.80E+01	2.34E+00	3.70E+00
o-cresol (2-hydroxytoluene)	tr	C 7 H 8 O 1	1.28E+01	2.55E+00	5.63E−01	−1.03E+02	4.98E+00	4.27E+03	−8.30E+03	−2.20E+02	2.10E+01	2.52E+00	2.08E+00
p-cresol (4-hydroxytoluene)	tr	C 7 H 8 O 1	1.28E+01	2.54E+00	5.45E−01	−1.03E+02	4.98E+00	4.05E+03	−8.16E+03	−2.20E+02	2.10E+01	2.52E+00	2.32E+00
2-aminotoluene	tr	C 7 H 9 N 1	1.39E+01	2.62E+00	5.87E−01	−1.08E+02	4.39E+00	4.12E+03	−8.27E+03	−2.27E+02	2.10E+01	2.52E+00	2.83E+00
ethyl benzene	tr	C 8 H 10	1.51E+01	2.97E+00	7.14E−01	−1.14E+02	6.50E−01	4.17E+03	−8.13E+03	−2.31E+02	2.10E+01	3.11E+00	3.54E+00
m-xylene	tr	C 8 H 10	1.53E+01	2.82E+00	8.07E−01	−1.14E+02	9.11E−01	4.18E+03	−8.14E+03	−2.32E+02	2.10E+01	2.52E+00	3.70E+00
4-ethenylcyclohexene	tr	C 8 H 12	1.71E+01	3.21E+00	1.03E+00	−1.20E+02	1.17E+00	4.65E+03	−9.08E+03	−2.40E+02	2.20E+01	3.11E+00	3.41E+00
cyclohexyl acetate	tr	C 8 H 14 O 2	2.04E+01	3.96E+00	1.42E+00	−1.38E+02	3.26E+00	7.27E+03	−1.46E+04	−2.89E+02	2.90E+01	4.00E+00	3.83E+00
di-butyl-ether	tr	C 8 H 18 O 1	2.47E+01	3.99E+00	5.95E−01	−1.45E+02	1.85E+00	6.30E+03	−1.23E+04	−2.89E+02	2.80E+01	8.00E+00	3.87E+00
acetophenone	tr	C 8 H 8 O 1	1.33E+01	2.86E+00	6.73E−01	−1.13E+02	8.57E+00	4.87E+03	−9.49E+03	−2.39E+02	2.30E+01	3.24E+00	2.91E+00
1,2,3-trimethylbenzene	tr	C 9 H 12	1.82E+01	3.24E+00	8.98E−01	−1.31E+02	9.79E−01	5.33E+03	−1.04E+04	−2.65E+02	2.40E+01	2.72E+00	3.70E+00
2,6-dimethyl-4-heptanone	tr	C 9 H 18 O 1	2.55E+01	4.04E+00	9.94E−01	−1.55E+02	4.90E+00	7.47E+03	−1.46E+04	−3.11E+02	3.00E+01	4.76E+00	3.76E+00
2-nonanone	tr	C 9 H 18 O 1	2.52E+01	4.26E+00	9.77E−01	−1.55E+02	4.82E+00	6.98E+03	−1.37E+04	−3.12E+02	3.00E+01	7.11E+00	3.51E+00
5-nonanone	tr	C 9 H 18 O 1	2.52E+01	4.33E+00	8.73E−01	−1.55E+02	4.82E+00	6.90E+03	−1.39E+04	−3.11E+02	3.00E+01	7.11E+00	3.00E+00
1-octadecanol	te	C 18 H 38 O 1	5.15E+01	9.02E+00	2.66E+00	−3.11E+02	2.32E+00	1.77E+04	−3.55E+04	−6.18E+02	5.80E+01	1.80E+01	4.30E+00
hexachloroethane	te	C 2 Cl 6	7.79E+00	3.65E+00	0.00E+00	−3.27E+01	4.58E+00	5.34E+03	−1.06E+04	−7.28E+01	2.50E+01	1.75E+00	3.65E+00
1,1,1,2-tetrachloroethane	te	C 2 H 2 Cl 4	7.74E+00	2.85E+00	0.00E+00	−3.49E+01	5.76E+00	3.17E+03	−6.25E+03	−7.32E+01	1.90E+01	1.63E+00	2.90E+00
1,1-dichloroethane	te	C 2 H 4 Cl 2	7.84E+00	1.88E+00	0.00E+00	−3.72E+01	4.44E+00	1.59E+03	−3.09E+03	−7.37E+01	1.30E+01	1.33E+00	2.86E+00
2-nitropropane	te	C 3 H 7 N 1 O 2	1.08E+01	1.74E+00	0.00E+00	−7.19E+01	1.20E+01	3.34E+03	−6.44E+03	−1.60E+02	1.80E+01	2.22E+00	2.86E+00
2-propanol	te	C 3 H 8 O 1	1.10E+01	1.41E+00	0.00E+00	−6.15E+01	6.12E+00	1.76E+03	−3.56E+03	−1.26E+02	1.30E+01	1.33E+00	3.70E+00
2-butanone	te	C 4 H 8 O 1	1.16E+01	1.76E+00	0.00E+00	−7.18E+01	9.01E+00	2.31E+03	−4.47E+03	−1.48E+02	1.50E+01	2.25E+00	3.44E+00
methyl propanoate	te	C 4 H 8 O 2	1.20E+01	1.88E+00	1.44E−01	−7.72E+01	6.04E+00	3.24E+03	−6.29E+03	−1.68E+02	2.20E+01	3.20E+00	3.70E+00
3-pentanone	te	C 5 H 10 O 1	1.43E+01	2.33E+00	2.50E−01	−8.85E+01	7.65E+00	3.17E+03	−6.15E+03	−1.81E+02	1.80E+01	3.20E+00	3.33E+00
n-propyl acetate	te	C 5 H 10 O 2	1.47E+01	2.40E+00	2.46E−01	−9.39E+01	4.14E+00	4.13E+03	−8.03E+03	−2.00E+02	2.10E+01	4.17E+00	3.97E+00
3-hexanone	te	C 6 H 12 O 1	1.70E+01	2.83E+00	4.56E−01	−1.05E+02	6.70E+00	4.06E+03	−7.92E+03	−2.13E+02	2.10E+01	4.17E+00	3.53E+00
cyclohexanol	te	C 6 H 12 O 1	1.66E+01	3.07E+00	1.08E+00	−1.05E+02	4.04E+00	4.40E+03	−8.59E+03	−2.16E+02	2.10E+01	2.34E+00	3.31E+00
isobutyl acetate	te	C 6 H 12 O 2	1.76E+01	2.76E+00	2.84E−01	−1.11E+02	3.67E+00	5.24E+03	−1.02E+04	−2.33E+02	2.40E+01	3.94E+00	4.13E+00
1,4-dichlorobenzene	te	C 6 H 4 Cl 2	9.57E+00	2.95E+00	6.81E−01	−7.88E+01	5.35E+00	3.94E+03	−7.70E+03	−1.67E+02	2.10E+01	2.52E+00	2.70E+00
3-methyl pyridine	te	C 6 H 7 N 1	1.13E+01	2.26E+00	4.48E−01	−9.09E+01	7.32E+00	3.24E+03	−6.29E+03	−1.91E+02	1.80E+01	2.34E+00	2.60E+00
2,4-dimethyl-3-pentanone	te	C 7 H 14 O 1	2.01E+01	3.09E+00	6.67E−01	−1.22E+02	5.57E+00	5.21E+03	−1.05E+04	−2.46E+02	2.40E+01	3.11E+00	3.55E+00
2-nitrotoluene	te	C 7 H 7 N 1 O 2	1.26E+01	2.88E+00	7.52E−01	−1.13E+02	7.89E+00	6.22E+03	−1.21E+04	−2.52E+02	2.60E+01	3.41E+00	2.95E+00
m-cresol (3-hydroxytoluene)	te	C 7 H 8 O 1	1.28E+01	2.54E+00	6.28E−01	−1.03E+02	4.98E+00	4.21E+03	−8.19E+03	−2.20E+02	2.10E+01	2.52E+00	2.38E+00
styrene	te	C 8 H 8	1.27E+01	2.61E+00	5.89E−01	−1.07E+02	1.20E+00	3.81E+03	−7.42E+03	−2.20E+02	2.00E+01	3.11E+00	3.70E+00
1,2,4-trimethyl-benzene	te	C 9 H 12	1.82E+01	3.24E+00	8.91E−01	−1.31E+02	9.79E−01	5.26E+03	−1.03E+04	−2.66E+02	2.40E+01	2.72E+00	3.70E+00

[a] χ = valence connectivity index; EE = exchange energy; HP = Hansen polarity; NNR = nuclear−nuclear repulsion; ENA = electron−nuclear attraction; RE = resonance energy; NFL = number of filled levels; κ = kappa index.

The simplest neighboorhood function is the bubble function, which is constant over the whole neighborhood of the winner neuron (node) and zero elsewhere. A more convenient function is the Gaussian neighborhood function defined by

$$\eta_o = \exp\left(\frac{-||r_c - r||^2}{2\sigma^2(t)}\right) \quad (4)$$

where $\sigma(t)$ is the neighborhood radius at t, which self-adapts after each epoch. The type of neighborhood function and the number of neurons used determine the sensitivity and the granularity of the map, respectively.

The learning rate $\alpha(t)$ in eq 3 is a decreasing function of t over [0,1]. A power series function in commonly used

$$\alpha(t) = \alpha_o\left(\frac{\alpha_T}{\alpha_o}\right)^{t/T} \quad (5)$$

where α_o and α_T are respectively the initial and final learning rates and T is the size of the training set, i.e., the number of epochs selected for training. Training is coarse over the first $T = 100$ epochs. During this initial training $\alpha(t)$ decreases monotonically according to eq 5, with $\alpha_o = 0.5$, and the neighborhood radius changes linearly between an 3 and 1. This coarse training is refined afterward for the following $T = 10\,000$ epochs, keeping the radius at 1 while decreasing $\alpha(t)$ according to 1/(training samples − 1), with $\alpha_o = 0.05$.

SOM is suitable for multivariable data analysis because of its prominent visualization capabilities. A preliminary idea of the number of clusters in the SOM as well as of their spatial relationships can be acquired by visual inspection of the map. The most common method used to visualize the cluster structure of SOM is the distance matrix or U-matrix. The U-matrix indicates the overall shape of the data set by means of the distances between prototype vectors of neighboring map units. Since neighbor nodes typically have similar prototype vectors, the U-matrix is closely related to a single linkage measure.[10] From the distance matrix, the different component maps for each descriptor and for the target activity variable can be obtained and clustered according to their topology. The main assumption of the current approach is that descriptors from the same cluster contribute with the same type of information to the QSARs. Thus, the repetitive inclusion of indices from the same cluster into a best set of molecular descriptors can only be justified after all relevant information from the other clusters has been considered and if their correlation with the target variable is higher than the average for the whole pool of descriptors.

To determine when all relevant information from other clusters has been considered, it is necessary to analyze the sensitivity of the maps to input variations. This sensitivity can be estimated by means of the dissimilarity between two maps L and M, measured as the averaged difference in their representation of the sample vectors used for training

$$D(L,M) = E\left[\frac{d_L(x) - d_M(x)}{d_L(x) + d_M(x)}\right] \quad (6)$$

In this equation E is the average expectation, and $d(x)$ the distance from x to the second BMU, denoted by $m_{c'(x)}$, beginning at the first BMU or winner neuron, denoted by $m_{c(x)}$. Of all possible paths between $m_{c(x)}$ and $m_{c'(x)}$ the shortest path passing continuously between neighbor units is selected

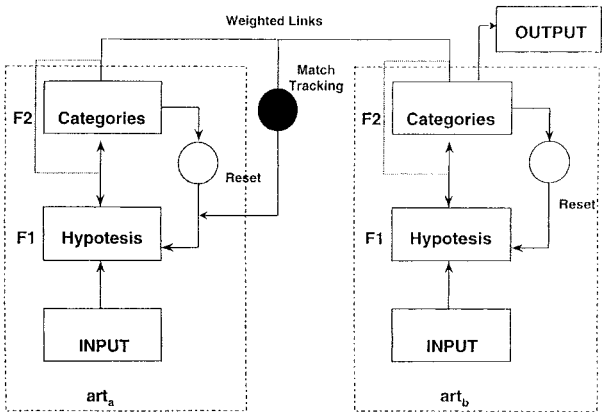

Figure 1. Block diagram of the fuzzy ARTMAP architecture.

$$d(x) = ||x - m_c(x)|| + \min_i \sum_{k=0}^{K_{c'(x)}-1} ||m_{Ii}(k) - m_{Ii}(k+1)|| \quad (7)$$

This distance, which was proposed by Kaski and Lagus,[18] combines an indication of the continuity of the mapping from the data set to the two-dimensional neural grid with a measure of the accuracy of the map in representing the data set.

The smallest average dissimilarity value calculated with eq 6 for any given set of descriptors indicates the similarity in quality and quantity of the information represented by the maps. Thus, the process of including indices to form the best set of molecular descriptors can be stopped when the dissimilarity measure stabilizes, i.e., the maps for these different indices are very similar.

Nevertheless, the clustering of the data set must be quite accurate to properly answer the question, what *similar* means? Since the final objective is to predict toxicity values, the clusters are labeled according to the toxicity of its center of mass and the quality of the clustering is determined by the correlation within the cluster members (homogeneity). The indices chosen by its minimal average dissimilarity (6) provide a good representation of the clusters formed and should constitute the best set of molecular descriptors for QSPR/QSAR.

FUZZY ARTMAP

Fuzzy ARTMAP,[26−28] the neural system chosen to establish QSAR models once the best set of molecular descriptors has been determined, is a supervised classifier that learns to categorize inputs as they are presented online using fuzzy logic to pattern recognize features. The architecture consists of a pair of fuzzy ART classifiers (art_A and art_B) that create stable recognition categories in response to arbitrary sequences of input patterns, as illustrated in Figure 1. The input vectors include both the molecular descriptors or target variable and the corresponding conjugates (complement coding). During supervised learning, the molecular descriptors (input patterns) of each chemical are presented to art_A, while the corresponding values for the target activity or property are presented to art_B. An associative learning network and an internal controller that ensure autonomous operation in real time link the information categorized by

these two modules. The controller is designed to create the minimal number of art_A categories, or hidden units, needed to match the accuracy criteria. It incorporates a minimum−maximum learning rule that enables fuzzy ARTMAP to learn quickly while ensuring minimum predictive error with maximum generalization. This scheme automatically links predictive success to category size on a trial by trail basis using only local operations. It works by increasing the vigilance parameter ρ_a of art_A by the minimal amount needed to correct the predictive error at art_B.

When an input vector $\vec{a}$ formed by the best set of descriptors is presented to the art_A module, the bottom-up activation from F_1^a causes the F_2^a layer to choose a category based on the fuzzy membership of the input in that category. Information of chosen category is then sent back to the F_1^a layer, and it is compared with the input vector $\vec{a}$. The fuzzy intersection of this top-down activation with the input vector produces a match value that indicates the confidence of the classification in that given category. The vigilance parameter ρ_a sets the threshold confidence value above which art_A accepts the category activated by an input as appropriate, rather than continue the search for a better class through an automatically controlled process of hypothesis testing. Module art_B follows an equivalent and simultaneous classification procedure when presented with the corresponding activity or property values during training.

The original fuzzy ARTMAP system was not designed to include predictive capabilities. The modification in architecture proposed by Giralt et al.[29] was implemented to allow predictions. Once the training of the neural system was completed, the input and hypothesis layers of art_B were disconnected so that predictions of the target variable (output) could be obtained from the category layer of art_B for any set of descriptors $\vec{a}$ presented as input to art_A.

METHODOLOGY

The general procedure proposed here to build QSPR/QSAR models is summarized in the flow diagram shown in Figure 2. Briefly, the experimental data for the two different sets of compounds and toxicities were processed to obtain the corresponding molecular structure using commercial software.[24] The geometry was optimized using MOPAC 6.0 and SOM was applied. A global map for all available molecular descriptors and target activity was calculated to select the best set of descriptors from the pool of available indices listed previously. The maps corresponding to each weight, i.e., the component planes (C-planes) for each input variable, were clustered based on visual inspection and by using linear correlations and curvilinear component analysis to unambiguously obtain an ordered representation of the component planes. This ordering implies that nearby maps have similar projections to the input data or molecular descriptors, i.e., highly correlated component planes depict similar spatial patterns over the two-dimensional neuronal distribution. The consistency of this clustering process for the component maps was checked against the values of the covariances between all variables calculated with the Pearson algorithm.

To ensure the best description of the input and output spaces, information from all clusters was first accounted for. To this end the descriptors with the highest covariance within

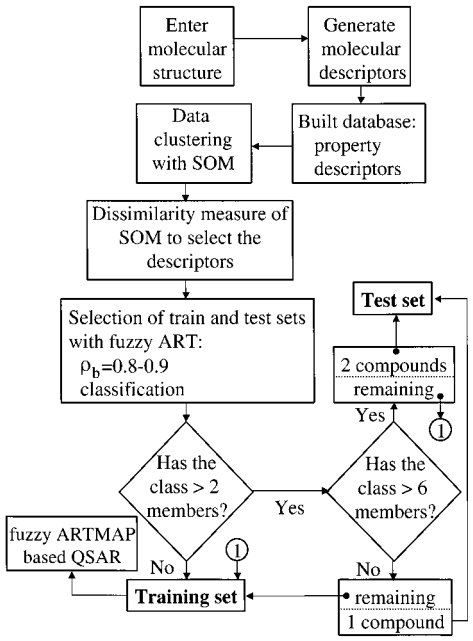

Figure 2. Flow diagram of the methodology followed to select the training and test sets.

each cluster were first selected, with the only restriction that the covariance was higher that the average value calculated for the whole pool of indices. After all relevant and nonrepetitive information from all clusters had been considered, with one descriptor per cluster included as the first elements of the best set, additional indices were added in order of decreasing absolute covariance with the target variable irrespectively of cluster membership. A map was obtained for each subset of descriptors formed. The dissimilarity between these maps was then calculated with eq 6 to obtain an unbiased measure of the information gained by the addition of each descriptor to the previous subset. The subset of descriptors whose map yielded the minimal average dissimilarity was selected as the best molecular input information to establish sound QSARs for the target variable considered. This is justified by the fact that any increase in dissimilarity indicates that the inclusion of the additional input information to the previous subset does not provide supplementary information to characterize de data set.

Once the best set of descriptors was chosen, the predictive fuzzy ARTMAP neural system was used to build the QSARs, following the procedure proposed by Espinosa et al.[16] First, the experimental toxicity data and the corresponding molecular descriptors of both sets of chemicals were assigned to either a train or a test subset using the fuzzy ART neural classifier. Assignment by classification is usually better than by a random selection procedure because the target property or activity might be unevenly distributed over the entire data set and the use of most of the relevant and redundant information during training is crucial for developing any neural network based QSPR/QSAR. The histograms for the 69 data of $-\log(LC_{50})$ and 155 of $\log(LD_{50})$, respectively, depicted in Figures 3 and 4 show that the distributions of both sets of toxicity values are indeed uneven over the range studied. Tables 1 and 2 show that about 85% of compounds from each set, 59 for LC_{50} and 135 for LD_{50}, were selected for training (tr) after presenting the input vectors formed by the molecular descriptors and the target toxicity value to

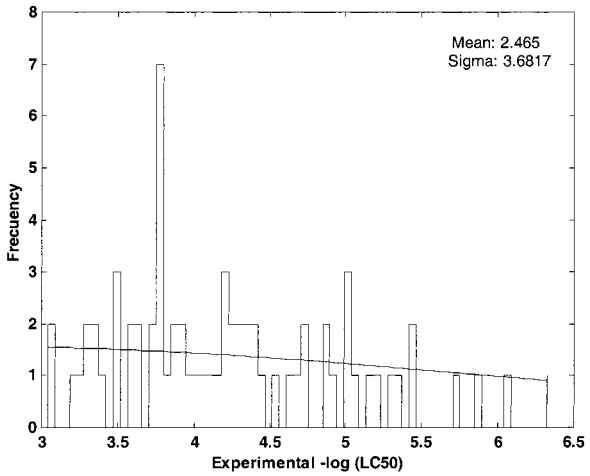

Figure 3. Histogram of $-\log(LC_{50})$ for the entire set of benzene derivatives.

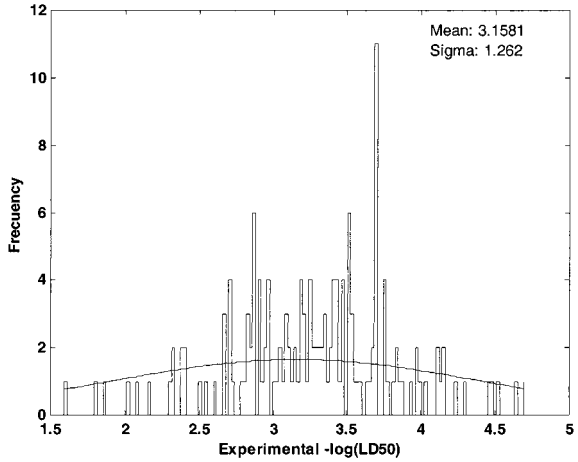

Figure 4. Histogram of $\log(LD_{50})$ for the entire heterogeneous set of organic compounds.

fuzzy ART. The two fuzzy ARTMAP based QSARs for toxicity of benzene derivatives and for the heterogeneous set of organic compounds were finally built from the training sets. It was checked that the above selection procedure yielded always the lowest errors in the prediction of the test data (te) for the two toxicity sets compared to the selection of a training set with random partitioning.

LC_{50} FOR BENZENE DERIVATIVES

Best Set of Indices. The LC_{50} toxicity data of 69 benzene derivatives were clustered with SOM, according to the integrated methodology summarized in Figure 2 and described above. Figure 5 depicts the distribution of six families (different kind of substituents on the aromatic ring) of benzene derivatives over the component plane for the target LC_{50}. The clusters of the different families are identified in this figure by capital letters: (A) halogen substituents; (B) hydroxyl; (C) nitro; (D) combined halogens and hydroxyl groups; (E) alkyl; and (F) additional combination of the previous ones. The derivatives are distributed according to family and molecular similarity, i.e., similar families are located nearby at positions where compounds are similar. For example, the family A, which is formed by halogenated derivatives, interfaces at several positions with family D, which is integrated by derivatives containing a combination of halogen and hydroxyl groups.

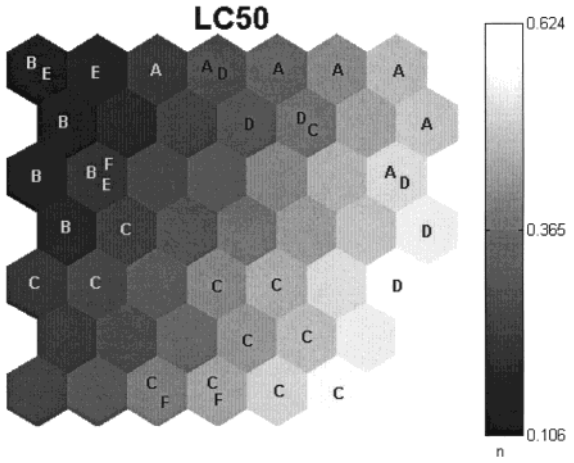

Figure 5. Overview of the distribution of the six families of benzene derivatives with (A) halogen, (B) hydroxyl, (C) nitro, (D) halogen and hydroxyl, (E) alkyl, and (F) additional substituents generated by SOM. The gray levels indicate the clustering intensity, *n*, of the LC_{50} data set.

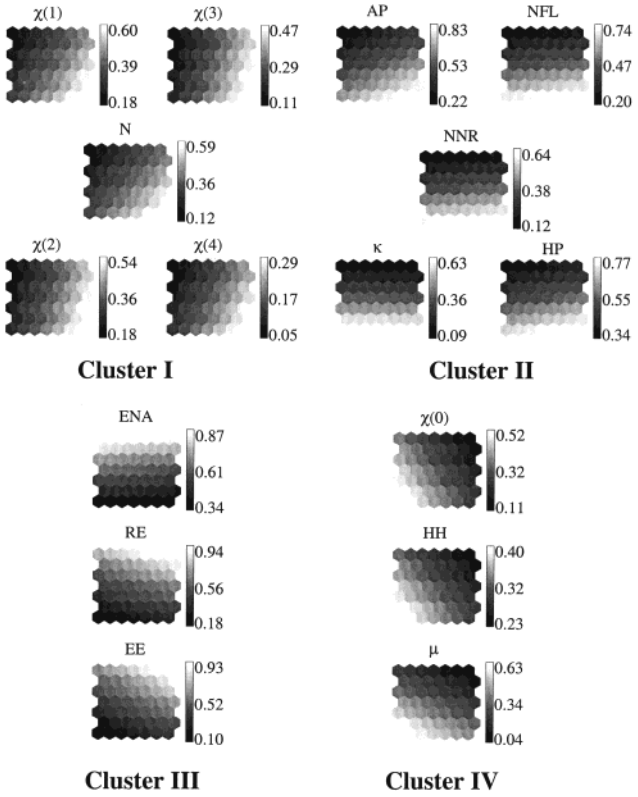

Figure 6. Clusters of the component maps for LC_{50} with current descriptors. The gray levels indicate distances.

Figure 6 shows the clusters of the component planes obtained by applying SOM to the target variable and the pool of 16 topological and quantum descriptors considered in the current study. Note that the component plane for LC_{50}, which has been considered in the clusterization process and is given in Figure 5, is not included in Figure 6 because is the target variable, i.e., the output information. The corresponding covariance matrix is included in Table 3. The analysis of these results for this homogeneous family of compounds shows unambiguously that

(i) The molecular connectivity indices of order one, two, three, and four and the sum of atomic numbers constitute the first cluster of indices. They show in Table 3

Table 3. Clustering and Covariance Matrix Using SOM for the LC_{50} of Benzene Derivatives[a]

ID		$^1\chi$	$^3\chi$	N	$^2\chi$	$^4\chi$	AP	NFL	NNR	κ	HP	ENA	RE	EE	$^0\chi$	HH	μ	LC_{50}
	$^1\chi$	1.000	0.927	0.911	0.968	0.615	0.798	0.544	0.489	0.351	0.386	−0.487	−0.143	−0.067	0.044	−0.062	0.013	0.688
	$^3\chi$	0.927	1.000	0.801	0.924	0.529	0.649	0.350	0.293	0.134	0.230	−0.290	0.071	0.134	−0.057	−0.040	−0.113	0.672
I	N	0.911	0.801	1.000	0.832	0.605	0.852	0.688	0.634	0.541	0.654	−0.634	−0.280	−0.144	−0.113	0.076	0.204	0.669
	$^2\chi$	0.968	0.924	0.832	1.000	0.608	0.659	0.350	0.288	0.131	0.217	−0.287	0.055	0.110	−0.001	−0.083	−0.113	0.645
	$^4\chi$	0.615	0.529	0.605	0.608	1.000	0.426	0.288	0.255	0.179	0.254	−0.255	−0.027	0.033	−0.053	0.021	0.124	0.400
	AP	0.798	0.649	0.852	0.659	0.426	1.000	0.906	0.875	0.782	0.739	−0.874	−0.642	−0.541	0.199	0.087	0.395	0.586
	NFL	0.544	0.350	0.688	0.350	0.288	0.906	1.000	0.994	0.964	0.825	−0.994	−0.855	−0.744	0.211	0.109	0.535	0.478
II	NNR	0.489	0.293	0.634	0.288	0.255	0.875	0.994	1.000	0.972	0.801	−1.000	−0.880	−0.773	0.227	0.100	0.533	0.454
	κ	0.351	0.134	0.541	0.131	0.179	0.782	0.964	0.972	1.000	0.813	−0.973	−0.900	−0.793	0.196	0.052	0.577	0.361
	HP	0.386	0.230	0.654	0.217	0.254	0.739	0.825	0.801	0.813	1.000	−0.804	−0.626	−0.461	−0.140	0.322	0.563	0.344
	ENA	−0.487	−0.290	−0.634	−0.287	−0.255	−0.874	−0.994	−1.000	−0.973	−0.804	1.000	0.880	0.773	−0.225	−0.103	−0.535	−0.453
III	RE	−0.143	0.071	−0.280	0.055	−0.027	−0.642	−0.855	−0.880	−0.900	−0.626	0.880	1.000	0.975	−0.543	−0.162	−0.625	−0.080
	EE	−0.067	0.134	−0.144	0.110	0.033	−0.541	−0.744	−0.773	−0.793	−0.461	0.773	0.975	1.000	−0.705	−0.128	−0.585	0.038
	$^0\chi$	0.044	−0.057	−0.113	−0.001	−0.053	0.199	0.211	0.227	0.196	−0.140	−0.225	−0.543	−0.705	1.000	0.105	0.228	−0.308
IV	HH	−0.062	−0.040	0.076	−0.083	0.021	0.087	0.109	0.100	0.052	0.322	−0.103	−0.162	−0.128	0.105	1.000	0.187	−0.268
	μ	0.013	−0.113	0.204	−0.113	0.124	0.395	0.535	0.533	0.577	0.563	−0.535	−0.625	−0.585	0.228	0.187	1.000	−0.092
	LC_{50}	0.688	0.672	0.669	0.645	0.400	0.586	0.478	0.454	0.361	0.344	−0.453	−0.080	0.038	−0.308	−0.268	−0.092	1.000

[a] $^{0-4}\chi$ = valence connectivity index; N = sum of atomic numbers; AP = average polarizability (PM3); NFL = number of filled levels; NNR = nuclear−nuclear repulsion; κ = kappa index; HP = Hansen polarity; ENA = electron−nuclear attraction; RE = resonance energy; EE = exchange energy; HH = Hansen hydrogen; μ = dipole moment.

consistent high covariances among themselves.

(ii) Both types of polarizability indices, the one calculated by the additive contribution of Hansen groups and the average one determined by a semiempirical calculation, are highly correlated with the number of filled levels, the nuclear−nuclear repulsion, and the kappa index. They all form the second cluster.

(iii) The electron nuclear attraction, the resonance energy, and the exchange energy are correlated among them and clustered together in the third group of indices.

(iv) The connectivity index of order zero, the dipole moment, and the Hansen hydrogen index are combined in the last cluster.

Table 4 includes the dissimilarities measures between 13 sets of molecular descriptors formed according to the methodology described before. The first set is formed by the representatives of the four clusters and the other 12 by adding the remaining 12 indices. The representatives of the four clusters given in Figure 5 and Table 3 are the connectivities of order zero and one ($^0\chi$, $^1\chi$), the average polarizability (AP), and the electron nuclear attraction (ENA). The average dissimilarity reaches the minimum value of 0.1387 when in addition to these four cluster representative indices the following six ones are included in order of decreasing absolute covariance with the target LC_{50}: the connectivity of order three ($^3\chi$), the sum of atomic numbers (N), the connectivity of order two ($^2\chi$), the number of filled levels (NFL), the nuclear−nuclear repulsion (NNR), and the connectivity of order four ($^4\chi$). If the best set of indices had been formed with the solely criteria of decreasing absolute covariance the only change would have been the inclusion of the kappa index (κ) instead of the connectivity of order zero ($^0\chi$). This apparently small modification causes small but very relevant changes in the classification of several chemicals as is discussed in the following subsection.

QSAR. The fuzzy ARTMAP model with the best set of descriptors was trained with 59 compounds selected with the fuzzy ART classifier (identified by *tr* in Table 1) with the vigilance parameter set to $\rho_a = 0.9$ and tested for the 10 chemicals identified with *te* in the same table. The −log(LC_{50}) predictions obtained with the fuzzy ARTMAP-based QSAR are depicted in Figure 7. This QSAR model predicts the −log(LC_{50}) of the complete data set of 69 compounds with an average absolute error of 0.02 log units (0.46%) and a standard deviation of 0.06 log units (1.35%). The average absolute error and standard deviation for the test set are 0.14 log units (3.18%) and 0.11 log units (2.04%), respectively.

Figure 7 also includes the predictions obtained with a fuzzy ARTAMP model based on the set of 10 descriptors formed by strict order of covariance, i.e., by substituting $^0\chi$ in the best set by κ. This change modifies the classification of the 3-methyl-2,4-dinitroaniline and 2,6-dimethylphenol in the test set, which are respectively identified in Figure 7 by the numbers 1 and 2. While the impact of these changes in classification on the overall performance is very small, the consequences are important from the point of view of individual errors or predictive reliability, as illustrated in Figure 7. Recognition categories obtained with the best set of descriptors do not show any misclassification. For example, the largest relative predictive error of 8.1% corresponds to *p*-cresol that was classified into the cluster of its homologous *m*-cresol.

The performance of the current fuzzy ARTMAP/QSAR is significantly superior than that for previously reported multilinear regression models (MLR)[19,20] as illustrated also in Figure 7. The absolute errors and standard deviations for predicted LC_{50} with the Hall et al.[19] and Gute and Basak[20] structure-toxicity models are similar and about 0.22 (5.2%) and 0.2 (4.3%) log units, compared to and 0.02 (0.05%) and 0.06 (1.4%) log units for the current model. A direct comparison of the largest relatives errors also corroborates the superior performance of the present fuzzy ARTMAP-based QSAR. While the largest relative error of the fuzzy ARTMAP model is 7.2% for 1,2,4,5-tetrachlorobenzene, the highest one for the MLR model reported by Gute and Basak[20] is 21.1% for the 5-methyl-2,4-dinitroaniline.

It is informative to examine the performance of current and previous models by inspecting the positional influence of functional group-specific errors given in Table 5 and Figure 8. The influence of ring position on toxicity proposed by Hall et al.[19] is presented in Table 5 for chlorobenzenes, halogenated derivatives with hydroxil substituents, alkyl

Table 4. Dissimilarity Measures between the Maps for the LC_{50} Set of Benzene Derivatives[a]

molecular descriptors used	$^0\chi$, $^1\chi$, AP, ENA	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP, HH	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP, HH, μ	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP, HH, μ, RE	$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP, HH, μ, RE, EE	av
$^0\chi$, $^1\chi$, AP, ENA		0.0484	0.0643	0.0964	0.1258	0.1620	0.1841	0.2192	0.2491	0.3386	0.3850	0.3998	0.4148	0.2240
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$	0.0484		0.0193	0.0522	0.0860	0.1229	0.1459	0.1825	0.2126	0.3022	0.3500	0.3652	0.3806	0.1890
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N	0.0643	0.0193		0.0342	0.0693	0.1089	0.1313	0.1688	0.1983	0.2866	0.3351	0.3504	0.3660	0.1777
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$	0.0964	0.0522	0.0342		0.0471	0.0870	0.1070	0.1474	0.1737	0.2590	0.3081	0.3237	0.3396	0.1646
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL	0.1258	0.0860	0.0693	0.0471		0.0460	0.0660	0.1093	0.1349	0.2246	0.2750	0.2908	0.3071	0.1485
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR	0.1620	0.1229	0.1089	0.0870	0.0460		0.0263	0.0695	0.0956	0.1913	0.2429	0.2592	0.2761	0.1406
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$	0.1841	0.1459	0.1313	0.1070	0.0660	0.0263		0.0498	0.0736	0.1685	0.2208	0.2372	0.2541	0.1387
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ	0.2192	0.1825	0.1688	0.1474	0.1093	0.0695	0.0498		0.0344	0.1319	0.1847	0.2013	0.2187	0.1431
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP	0.2491	0.2126	0.1983	0.1737	0.1349	0.0956	0.0736	0.0344		0.1031	0.1549	0.1717	0.1894	0.1493
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP, HH	0.3386	0.3022	0.2866	0.2590	0.2246	0.1913	0.1685	0.1319	0.1031		0.0562	0.0733	0.0935	0.1857
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP, HH, μ	0.3850	0.3500	0.3351	0.3081	0.2750	0.2429	0.2208	0.1847	0.1549	0.0562		0.0224	0.0437	0.2149
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP, HH, μ, RE	0.3998	0.3652	0.3504	0.3237	0.2908	0.2592	0.2372	0.2013	0.1717	0.0733	0.0224		0.0242	0.2266
$^0\chi$, $^1\chi$, AP, ENA, $^3\chi$, N, $^2\chi$, NFL, NNR, $^4\chi$, κ, HP, HH, μ, RE, EE	0.4148	0.3806	0.3660	0.3396	0.3071	0.2761	0.2541	0.2187	0.1894	0.0935	0.0437	0.0242		0.2423

[a] $^{0-4}\chi$ = valence connectivity index; κ = kappa index; N = sum of atomic numbers; HP = Hansen polarity; HH = Hansen hydrogen; NFL = number of filled levels; μ = dipole moment; AP = average polarizability (PM3); RE = resonance energy; EE = exchange energy; ENA = electron−nuclear attraction; NNR = nuclear−nuclear repulsion.

benzenes, and mixed phenol derivatives. For the chlorobenzene family the absolute mean errors for the Gute and Basak[20] and Hall et al.[19] models are 0.11 (2.2%) and 0.04 (1%) log units, respectively, compared to 0.08 log units (1.6%) for the fuzzy ARTMAP model. The corresponding standard deviations are 0.11 (2.3%), 0.08 (1.7%), and 0.17 (3.2%) log units. The good performance of MLR models for the homogeneous chlorobenzene set contrasts with the relatively high errors of fuzzy ARTMAP, which has immense generalization capabilities when properly trained. Only 59 chemicals were used to train the fuzzy ARTMAP model for LC_{50}, seven of which belonging to the chlorobenzene family. In addition, the quantum chemical descriptors included in the present pool of 16 molecular indices may not be the best choice for providing the required information to distinguish chlorobenzenes. Inspection of Table 5 shows that indeed the 1,3-dichlorobenzene (*te*) cannot be distinguished from the 1,4-dichlorobenzene (*tr*), and the 1,2,3,4-tetrachlorobenzene (*tr*) from the 1,2,4,5-tetrachlorobenzene (*te*). In fact, fuzzy ARTMAP clearly outperforms MLR models and yields LC_{50} accurate predictions for the other three families of compounds in Table 5. The errors and standard deviations for all the compounds included in the training set are zero and

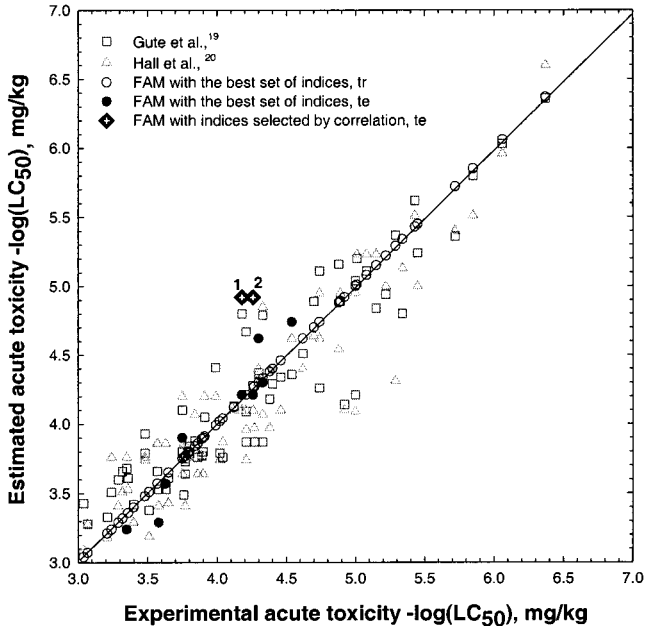

Figure 7. Comparison of experimental with predicted $-\log(LC_{50})$ toxicity values of benzene derivatives.

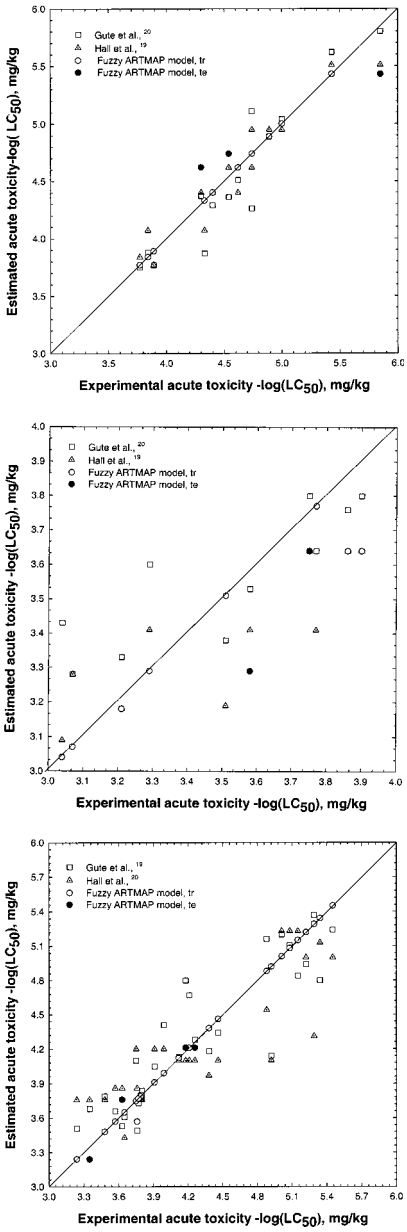

Figure 8. Comparison of the positional influence on the −log(LC_{50}) toxicity measured and predicted for the three families of benzene derivatives with (a) halogen, (b) hydroxyl, and (c) nitro substituents.

insignificant for the test compound 2,4,6-trichlorophenol.

Figures 8a−c depict the experimental and predicted LC_{50} for benzene derivatives with halogen, hydroxyl, and nitro substituents. The first family is an extension of the chlorobenzes included in Table 5. The absolute mean errors for predictions in Figure 8a are 0.16 (3.5%), 0.14 (3%), and 0.07 (1.4%) log units for the Gute and Basak[20] and Hall et al.[19] fuzzy ARTMAP models, respectively. The results for the hydroxyl substituents plotted in Figure 8b show that the fuzzy ARTMAP-based QSAR performs well with an absolute mean error 0.09 (2.4%) log units compared to 0.16 (4.8%) log units for the Gute and Basak[20] and 0.19 (5.2%) log units for the Hall et al.[19] models. Finally, Figure 8c illustrates the influence of nitro substituents on the toxicity LC_{50} values in a rather heterogeneous data set. In this family fuzzy ARTMAP outperforms literature MLR models[19,20] in terms of errors by a factor larger than ten.

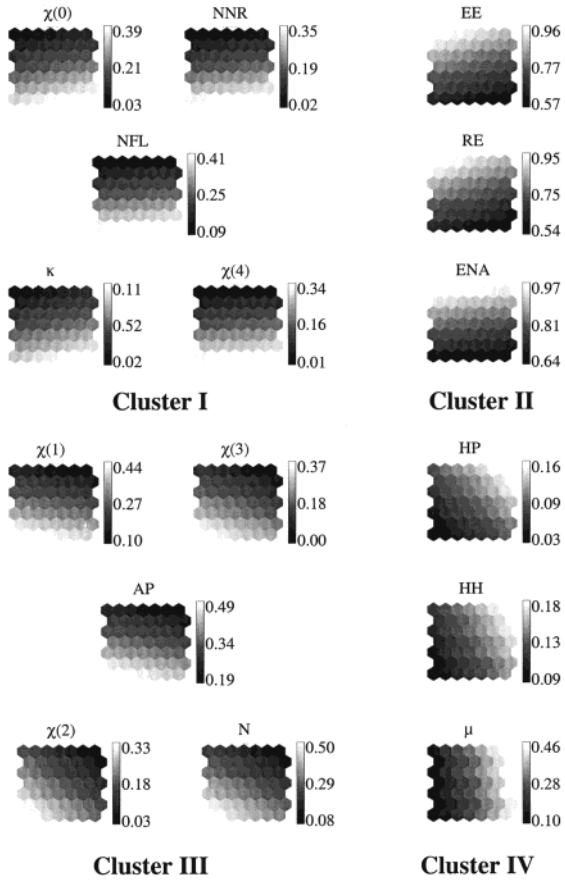

Figure 9. Clusters of the component maps for LD_{50} with current descriptors. The gray levels indicate distances.

The enormous differences between previous MLR models[19,20] and the current fuzzy ARTMAP-based QSAR cannot be attributed to differences in the quality and quantity of the molecular information used in the three cases but to the nonlinear nature and the superior performance of cognitive classifiers, as has been already demonstrated in the literature. Gute and Basak[20] used multiple linear regression analysis with seven parameters to model LC_{50}: the Zagreb group parameter, the path of length nine, structural information for the zeroth order neighborhood of vertices in a hydrogen-filled graph, the 3-D Wiener number for hydrogen-filled structures, the second lowest unoccupied molecular orbital (LUMO), the heat of formation, and the dipole moment. On the other hand, Hall et al.[19] used multiple linear regression analysis to obtain the coefficients in a Free-Wilson equation for each substituent group.

To distinguish the effect of molecular descriptors from that of algorithm in the performance of the current QSAR for LC_{50}, a fuzzy ARTMAP model using the indices of Gute and Basak[20] was also derived. This model yielded predictions for the complete data set of 69 benzene derivatives with an absolute mean error of 0.02 (0.58%), which is comparable to that of 0.02 (0.46%) for the current fuzzy ARTMAP model. Nevertheless, the generalization capability of the current model with the best set of descriptors is slightly superior, as shown by the predictions and errors listed in Table 6 for the 10 compounds that constitute the test set. The absolute mean error and standard deviation for the fuzzy ARTMAP-based QSAR with the indices of Gute and Basak[20] are 0.17 (3.84%) and 0.13 (2.67%), respectively, compared

Table 5. Positional Influence of Functional Groups on Acute Toxicity LC_{50}

name	formula	exp $-\log(LC_{50})$	Gute et al.[20]	Hall et al.[19]	FAM $-\log(LC_{50})$
		Chlorobenzenes			
chlorobenzene	C 6 H 5 Cl 1	3.77	3.75	3.84	3.77
1,2-dichlorobenzene	C 6 H 4 Cl 2	4.40	4.29	4.40	4.40
1,3-dichlorobenzene[a]	C 6 H 4 Cl 2	4.30	4.37	4.40	4.62
1,4-dichlorobenzene	C 6 H 4 Cl 2	4.62	4.51	4.40	4.62
1,2,3-trichlorobenzene	C 6 H 3 Cl 3	4.89	4.89	4.89	4.89
1,2,4-trichlorobenzene	C 6 H 3 Cl 3	5.00	5.04	5.00	5.00
1,3,5-trichlorobenzene	C 6 H 3 Cl 3	4.74	5.11	4.74	4.74
1,2,3,4-tetrachlorobenzene	C 6 H 2 Cl 4	5.43	5.62	5.43	5.43
1,2,4,5-tetrachlorobenzene[a]	C 6 H 2 Cl 4	5.85	5.80	5.85	5.43
		Halogen and Hydroxyl Substituents			
2,3,4,5,6-pentachlorophenol	C 6 H 1 O 1 Cl 5	6.06	6.03	5.96	6.06
2,3,4,5-tetrachlorophenol	C 6 H 2 O 1 Cl 4	5.72	5.36	5.40	5.72
2,4,6-tribromophenol	C 6 H 3 O 1 Br 3	4.70	4.89	4.64	4.70
2,4,6-trichlorophenol[a]	C 6 H 3 O 1 Cl 3	4.33	4.79	4.85	4.30
2,4-dichlorophenol	C 6 H 4 O 1 Cl 2	4.30	4.33	4.30	4.30
		Alkyl Substituents			
toluene	C 7 H 8	3.32	3.66	3.51	3.32
1,2-dimethylbenzene	C 8 H 10	3.48	3.93	3.74	3.48
1,4-dimethylbenzene	C 8 H 10	4.21	3.87	3.74	4.21
1,2,4-trimethylbenzene	C 9 H 12	4.21	4.09	3.96	4.21
benzene	C 6 H 6	3.40	3.42	3.29	3.40
		Halogenated, Nitro, and Hydroxyl Substituents			
4-chloro-3-methylphenol	C 7 H 7 O 1 Cl 1	4.27	3.87	3.97	4.27
2,4-dinitrophenol	C 6 H 4 N 2 O 5	4.04	3.76	3.87	4.04
2-methyl-4,6-dinitrophenol	C 7 H 6 N 2 O 5	5.00	4.21	4.09	5.00
4-nitrophenol	C 6 H 5 N 1 O 3	3.36	3.61	3.53	3.36
2-chlophenol	C 6 H 5 O 1 Cl 1	4.02	3.79	3.74	4.02

[a] Compounds used for testing.

Table 6. Comparison of the Performance of Fuzzy ARTMAP-Based QSARs for LC_{50} of Benzene Derivatives Using Gute and Basak[20] Descriptors and the Current Best Set during Generalization (Test Compounds)

name	formula	exp $-\log(LC_{50})$	Gute and Basak[20]	relative error %	current model	relative error
1,2,4,5-tetrachlorobenzene	C 6 H 2 Cl 4	5.85	5.43	7.18	5.43	7.18
2,4,6-trichlorophenol	C 6 H 3 O 1 Cl 3	4.33	4.30	0.69	4.30	0.69
1,3-dichlorobenzene	C 6 H 4 Cl 2	4.30	4.62	7.44	4.40	2.32
2,4-dichlorotoluene	C 7 H 6 Cl 2	4.54	4.74	4.40	4.74	4.40
3-nitrotoluene	C 7 H 7 N 1 O 2	3.63	3.75	3.35	3.57	1.65
5-methyl-2,6-dinitroaniline	C 7 H 7 N 3 O 4	4.18	4.21	0.72	4.21	0.72
2-methyl-5-nitroaniline	C 7 H 8 N 2 O 2	3.35	3.24	3.24	3.24	3.28
p-cresol (4-hydroxytoluene)	C 7 H 8 O 1	3.58	3.84	7.26	3.77	5.31
3-methyl-2,4-dinitroaniline	C 7 H 7 N 3 O 4	4.26	4.21	1.17	4.21	1.17
2,6-dimethylphenol	C 8 H 10 O 1	3.75	3.86	2.93	3.86	2.93

to 0.14 (3.18%) and 0.11 (2.04%) for the current model. These differences are due to changes in the classification of 1,3-dichlorobenzene, 3-nitrotoluene, and *p*-cresol. For example, the current model classifies correctly the *p*-cresol with its homologous *o*-cresol and predicts its toxicity with a 5.31% error, while the fuzzy ARTMAP model with the descriptors of Gute and Basak[20] misclassifies it with the 3-chlorotoluene and the predictive error increases slightly to 7.26%, as shown in Table 6.

LD_{50} FOR ORGANIC COMPOUNDS

Best Set of Indices. The clusters of component maps, which were obtained from the U-matrix calculated for the set of indices and the corresponding LD_{50} values given in Table 2, are depicted in Figure 9. The component maps in the four clusters are in good agreement with the Pearson covariances presented in Table 7. The analysis of these results for this heterogeneous family of compounds shows unambiguously that

(i) The molecular connectivity indices of order zero and four are correlated with the nuclear nuclear repulsion index, the number of filled levels, and the kappa index.

(ii) The exchange energy, the resonance energy, and the electron nuclear attraction indices are correlated among themselves.

(iii) The connectivity indices of order one, two, and three are correlated with the average polarizability and the sum of atomic numbers.

(iv) The Hansen polarizability, the Hansen hydrogen planes, and the dipole moment are correlated.

It is instructive to compare the clusters of the component maps for the heterogeneous LD_{50} set in Figure 9 (Table 7) with those for the homogeneous LC_{50} set in Figure 6 (Table 3). In the homogeneous set of benzene derivatives all topological information contained in the connectivity indices and in the sum of atomic numbers (mostly in cluster I of Figure 6) was very relevant in terms of both covariances with LC_{50} and cluster representatives (cluster IV). Also, the need to distinguish isomers gave an important roll to the average polarizability. The introduction of heterogeneity reduces the impact of any trend related to homogeneity and

Table 7. Clustering and Covariance Matrix Using SOM for the LD_{50} of a Heterogeneous Set of Organic Compounds[a]

ID		$^0\chi$	NNR	NFL	κ	$^4\chi$	EE	RE	ENA	$^1\chi$	$^3\chi$	AP	$^2\chi$	N	HP	HH	μ	LD_{50}
	$^0\chi$	1.000	0.881	0.874	0.900	0.799	−0.947	−0.933	−0.876	0.846	0.614	0.568	0.441	0.514	−0.308	−0.063	−0.049	0.432
	NNR	0.881	1.000	0.988	0.845	0.852	−0.900	−0.910	−0.984	0.886	0.751	0.724	0.537	0.684	−0.205	−0.063	−0.005	0.392
I	NFL	0.874	0.988	1.000	0.842	0.831	−0.900	−0.907	−0.974	0.891	0.754	0.756	0.559	0.705	−0.221	−0.072	−0.027	0.389
	κ	0.900	0.845	0.842	1.000	0.734	−0.859	−0.844	−0.848	0.815	0.573	0.518	0.374	0.531	−0.231	−0.026	−0.001	0.365
	$^4\chi$	0.799	0.852	0.831	0.734	1.000	−0.847	−0.850	−0.826	0.825	0.714	0.702	0.420	0.565	−0.282	−0.156	−0.114	0.356
	EE	−0.947	−0.900	−0.900	−0.859	−0.847	1.000	0.993	0.883	−0.788	−0.599	−0.653	−0.339	−0.448	0.276	0.079	−0.004	−0.402
II	RE	−0.933	−0.910	−0.907	−0.844	−0.850	0.993	1.000	0.891	−0.770	−0.593	−0.659	−0.323	−0.439	0.245	0.055	−0.057	−0.391
	ENA	−0.876	−0.984	−0.974	−0.848	−0.826	0.883	0.891	1.000	−0.875	−0.734	−0.688	−0.530	−0.677	0.190	0.046	−0.014	−0.388
	$^1\chi$	0.846	0.886	0.891	0.815	0.825	−0.788	−0.770	−0.875	1.000	0.842	0.735	0.752	0.861	−0.349	−0.181	−0.229	0.326
	$^3\chi$	0.614	0.751	0.754	0.573	0.714	−0.599	−0.593	−0.734	0.842	1.000	0.702	0.626	0.789	−0.267	−0.155	−0.259	0.220
III	AP	0.568	0.724	0.756	0.518	0.702	−0.653	−0.659	−0.688	0.735	0.702	1.000	0.571	0.684	−0.313	−0.232	−0.176	0.188
	2c	0.441	0.537	0.559	0.374	0.420	−0.339	−0.323	−0.530	0.752	0.626	0.571	1.000	0.912	−0.258	−0.165	−0.284	0.174
	N	**0.514**	0.684	0.705	0.531	0.565	−0.448	−0.439	−0.677	0.861	0.789	0.684	0.912	1.000	−0.224	−0.136	−0.245	0.167
	HP	−0.308	−0.205	−0.221	−0.231	−0.282	0.276	0.245	0.190	−0.349	−0.267	−0.313	−0.258	−0.224	1.000	0.658	0.472	−0.289
IV	HH	−0.063	−0.063	−0.072	−0.026	−0.156	0.079	0.055	0.046	−0.181	−0.155	−0.232	−0.165	−0.136	0.658	1.000	0.196	−0.205
	μ	−0.049	−0.005	−0.027	−0.001	−0.114	−0.004	−0.057	−0.014	−0.229	−0.259	−0.176	−0.284	−0.245	0.472	0.196	1.000	−0.083
	LD_{50}	0.432	0.392	0.389	0.365	0.356	−0.402	−0.391	−0.388	0.326	0.220	0.188	0.174	0.167	−0.289	−0.205	−0.083	1.000

[a] χ = valence connectivity index; NNR = nuclear−nuclear repulsion; NFL = number of filled levels; κ = kappa index; EE = exchange energy; RE = resonance energy; ENA = electron−nuclear attraction; AP = average polarizability (PM3); N = sum of atomic numbers; HP = Hansen polarity; HH = Hansen hydrogen; μ = dipole moment.

Table 8. Dissimilarity Measures between the Maps for the LD_{50} of the Heterogeneous Set of Compounds[a]

molecular descriptors used	$^0\chi$, $^1\chi$, EE, HP	$^0\chi$, $^1\chi$, EE, HP, NNR	$^0\chi$, $^1\chi$, EE, HP, NNR, RE	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA,κ, $^4\chi$, $^3\chi$	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA,κ, $^4\chi$, $^3\chi$, HH	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA,κ, $^4\chi$, $^3\chi$, HH, AP	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$, $^3\chi$, HH, $^2\chi$	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA κ, $^4\chi$, $^3\chi$, HH, $^2\chi$, N	$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$, $^3\chi$, HH, $^2\chi$, N, μ
$^0\chi$, $^1\chi$, EE, HP		0.0491	0.0845	0.1075	0.1265	0.1896	**0.2471**	0.2815	0.3221	0.3410	0.3615	0.3908	0.4499
$^0\chi$, $^1\chi$, EE, HP, NNR	0.0491		0.0390	0.0612	0.0806	0.1446	**0.2035**	0.2387	0.2802	0.2996	0.3208	0.3513	0.4136
$^0\chi$, $^1\chi$, EE, HP, NNR, RE	0.0845	0.0390		0.0260	0.0468	0.1095	**0.1686**	0.2049	0.2466	0.2664	0.2881	0.3194	0.3835
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL	0.1075	0.0612	0.0260		0.0219	0.0856	**0.1456**	0.1823	0.2241	0.2443	0.2663	0.2981	0.3631
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA	0.1265	0.0806	0.0468	0.0219		0.0664	**0.1269**	0.1641	0.2058	0.2262	0.2484	0.2806	0.3461
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ	0.1896	0.1446	0.1095	0.0856	0.0664		**0.0630**	0.1013	0.1431	0.1646	0.1881	0.2221	0.2881
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$	0.2471	0.2035	0.1686	0.1456	0.1269	0.0630		0.0421	0.0821	0.1049	0.1300	0.1638	0.2304
$^0\chi$, $^1\chi$ EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$, $^3\chi$	0.2815	0.2387	0.2049	0.1823	0.1641	0.1013	**0.0421**		0.0460	0.0671	0.0935	0.1290	0.1953
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$, $^3\chi$, HH	0.3221	0.2802	0.2466	0.2241	0.2058	0.1431	**0.0821**	0.0460		0.0296	0.0601	0.0913	0.1527
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$, $^3\chi$, HH, AP	0.3410	0.2996	0.2664	0.2443	0.2262	0.1646	**0.1049**	0.0671	0.0296		0.0348	0.0707	0.1340
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$, $^3\chi$, HH, $^2\chi$	0.3615	0.3208	0.2881	0.2663	0.2484	0.1881	**0.1300**	0.0935	0.0601	0.0348		0.0468	0.1122
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$, $^3\chi$, HH, $^2\chi$, N	0.3908	0.3513	0.3194	0.2981	0.2806	0.2221	**0.1638**	0.1290	0.0913	0.0707	0.0468		0.0770
$^0\chi$, $^1\chi$, EE, HP, NNR, RE, NFL, ENA, κ, $^4\chi$, $^3\chi$, HH, $^2\chi$, N, μ	0.4499	0.4136	0.3835	0.3631	0.3461	0.2881	**0.2304**	0.1953	0.1527	0.1340	0.1122	0.0770	
average	0.2459	0.2068	0.1819	0.1688	0.1617	0.1472	**0.1423**	0.1455	0.1570	0.1653	0.1792	0.2034	0.2621

[a] $^{0-4}\chi$ = valence connectivity index; κ = kappa index; N = sum of atomic numbers; HP = Hansen polarity; HH = Hansen hydrogen; NFL = number of filled levels; μ = dipole moment; AP = average polarizability (PM3); RE = resonance Energy; EE = exchange energy; ENA = electron−nuclear attraction; NNR = nuclear−nuclear repulsion.

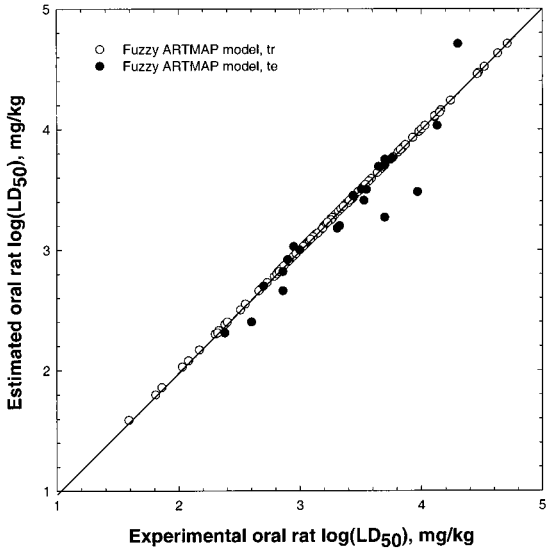

Figure 10. Comparison of experimental with predicted log(LD_{50}) toxicity values of an heterogeneous set of organic compounds.

the need for redundancy in topological information and enhances the roll of quantum chemical descriptors.

Again, maps for the different subsets of descriptors are compared by the dissimilarity measures given in Table 8. The best set of indices for LD_{50} is integrated by the following 10 molecular descriptors: the molecular connectivity indices of order zero, one, and four, the kappa index, the exchange energy, the Hansen polarizability, the nuclear nuclear repulsion, the resonance energy, the number of filled levels, and the electron nuclear attraction. The changes with respect to the LC_{50} best set are the substitution of the connectivity of second and third order, the sum of atomic numbers and the averatge polarizability by the exchange energy, the Hansen polarizability, the resonance energy, and the kappa index.

QSAR. The performance of the fuzzy ARTMAP toxicity model for LD_{50} with the best set of descriptors selected above is depicted in Figure 10, where predictions are compared with experimental values. The mean absolute error is 0.02 (0.53%) log units with a standard deviation of 0.07 (1.84%) log units for the entire data set listed in Table 2. The absolute mean errors for the training and test sets are 0.00 (0.06%) and 0.13 (3.68%) log units, respectively. The standard deviations for these two sets respectively are 0.07 (1.84%) and 0.15 (3.92%) log units. The largest individual error of 12.50% obtained with this model corresponds to the *n*-propyl acetate. This and other smaller errors are not caused by misclassification of the any of the 155 heterogeneous compounds included in the LD_{50} set. For example, the *n*-propyl acetate is correctly classified with its homologous isopropyl acetate. It should be noted that specific interactions play an essential roll in the toxicity of any compound; small stereochanges can alter significantly toxicity and, thus, generate predictive errors as large as the 12.50% above without misclassification.

CONCLUSIONS

The identification of relevant molecular indices for QSAR can be carried out systematically by the use of self-organizing maps, since it is possible to establish the influence of input parameters in their topology and to cluster them according to similarities. The selection of the minimum set of most significant indices necessary to distinguish, for example, between toxic and nontoxic compounds has been carried out by incorporating the more representative indices of each cluster as well as the ones with higher absolute covariance with the target variable. The two best sets with 10 different indices for LC_{50} and LD_{50} have been used as input to a fuzzy ARTMAP classifier, modified to effect predictive capabilities.

The fuzzy ARTMAP-based QSAR for LC_{50} predicts the toxicity of 69 benzene derivatives without misclassifications and with average absolute errors of 0.02 (0.46%) and 0.14 (3.18%) log units for the whole and test sets, respectively. This neural system outperforms the two previously reported QSAR models[19,20] both in terms of overall errors and classification. The fuzzy ARTMAP based QSAR for the toxicity LD_{50}, of 155 heterogeneous compounds yield predictions with mean errors of 0.02 (0.53%) log units for the complete data set and 0.13 (3.68%) log units for the test set. The largest single error of 12.50% observed corresponds to the *n*-propyl acetate and is not caused by misclassification. As a consequence, the proposed integrated SOM-fuzzy ARTMAP approach is a useful tool to establish systematically sound QSAR/QSPR models.

ACKNOWLEDGMENT

The authors are grateful for the financial support received from the "Dirección General de Investigación Científica y Técnica", projects no. PB96-1011 and PPQ2000-1339, and from the CIRIT "Programa de Grups de Recerca Consolidats de la Generalitat de Catalunya", projects no. 1998SGR-00102 and 2000SGR-00103. The paper was written while A.A. was a visiting scholar at the University of California Berkeley with the fellowship of the Spanish Ministry of Science and Education, PR-2001-0154. The authors are also very grateful to Dr. Gute, Prof. Basak, and Prof. Kier for providing valuable data and information about their previous studies concerning the prediction of toxicity LC_{50} of benzene derivatives.

REFERENCES AND NOTES

(1) Kier, L.; Hall, L. *Molecular Connectivity in Chemistry and Drug Research;* Academic Press: New York, 1976.
(2) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc*. **1975**, *97*, 6609.
(3) Randic, M.; Trinajstic, N. Comparative Structure−Property Studies: The Connectivity Basis. *J. Mol. Struct*. **1993**, *284*, 209.
(4) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc*. **1947**, *69*, 17.
(5) Stanton, D.; Egolf, L.; Jurs, P.; Hicks, M. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans and Thiophenes. *J. Chem. Inf. Comput. Sci*. **1991**, *31*, 301.
(6) Stanton, D.; Egolf, L.; Jurs, P.; Hicks M. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 306.
(7) Kohonen, T. Self-Organizing Formation of Topologically Correct Feature Maps. *Biological Cybernetics* **1982**, *43*, 59.
(8) Kohonen, T. Physiological Interpretation of the Self-organizing Map Algorithm. *Neural Networks* **1993**, *6*, 895.
(9) Kohonen, T. The Self-Organizing Map. *Proc. IEEE* **1990**, *78*, 1464.
(10) Vesanto, J. SOM-Based Data Visualization Methods. *Intelligent Data Analysis* **1999**, *6*, 111.
(11) Anzali, S.; Gasteiger, J.; Holzgrabe, U.; Polanski, J.; Sadowski, J.; Teckentrup, A.; Wagener, M. The Use of Self-Organizing Neural Networks in Drug Design. In *3D QSAR in Drug Design, 2;* Kubinyi, H., Folkers, G., Martin, Y. C., Ed.; Kluwer/ESCOM: Dordrecht, The Netherlands, 1998; p 273.

(12) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc*. **1995**, *117*, 7769.
(13) Gasteiger, J.; Li, X.; Rudolph, C. J.; Sadowski, J.; Zupan, J. Representation of Molecular Electrostatic Potentials by Topological Feature Maps. *J. Am. Chem. Soc*. **1994**, *116*, 4608.
(14) Gasteiger, J.; Li, X.; Uschold, A. The Beauty of Molecular Surfaces as Revealed by Self-Organizing Neural Networks. *J. Mol. Graphics* **1994**, *12*, 90.
(15) Espinosa, G.; Yaffe, D.; Cohen, Y.; Arenas, A.; Giralt, F. Neural Network Based Quantitative Structural Property Relations (QSPRs) for Predicting Boiling Points of Aliphatic Hydrocarbons. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 859.
(16) Espinosa, G.; Yaffe, D.; Arenas, A.; Cohen, Y.; Giralt, F. A Fuzzy ARTMAP Based Quantitative Structure−Property Relationships (QSPRs) for Predicting Physical Properties of Organic Compounds. *Ind. Eng. Chem. Res.* **2001**, *40*, 2757.
(17) Kaski, S.; Kohonen, T. Winner-take-all Networks for Physiological Models of Competitive Learning. *Neural Networks* **1994**, *7*, 973.
(18) Kaski, S.; Lagus, K. Comparing Self-organizing Maps. In *Proceedings of ICANN'96*; 1996; p 809.
(19) Hall, L.; Kier, L.; Phipps, G. Structure−Activity Relationship Studies on the Toxicities of Benzene Derivatives I an Additivity Model. *Environ. Toxicol. Chem*. **1984**, *3*, 355.
(20) Gute, B.; Basak, S. Predicting Acute Toxicity (LC_{50}) of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* **1997**, *7*, 117.
(21) Zhen, W.; Tropsha, A. Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 185.
(22) Ivanciuc, O.; Taraviras, S.; Cabrol-Bass, D. Quasi-orthogonal Basis Sets of Molecular Graph Descriptors as a Chemical Diversity Measure. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 126.
(23) Hansen, C. M. The Three-dimensional Solubility Parameter − Key to Paint Component Affinities. III. Independent Calculation of the Parameter Components. *J. Paint Technol*. **1967**, *39*, 511.
(24) Molecular Modeling Pro.− Revision 3.1; ChemSW Inc.: 1998.
(25) Stewart, J. L. MOPAC 6.0. Quantum Chemistry Program Exchange No. 455, Bloomington, IN, 1989.
(26) Carpenter, G.; Grossberg, S. A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine. *Comput. Vision Graphics, Image Processing* **1987**, 37, 54.
(27) Carpenter, G.; Grossberg, S.; Marcuson, N.; Reynolds, J.; Rosen, D. Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analogue Multidimensional Maps. *IEEE Trans. Neural Networks* **1992**, *3*, 698.
(28) Carpenter, G.; Grossberg, S.; Marcuzon, N.; Rosen, D. B. Fuzzy ART: Fast Stable Learning and Categorization of Analogue Patterns by an Adaptive Resonance System. *Neural Networks* **1991**, *4*, 759.
(29) Giralt, F.; Arenas, A.; Ferre-Giné, J.; Rallo R. The Simulation and Interpretation of Turbulence with a Cognitive Neural System. *Phys. Fluids* **2000**, *12*, 1826.

CI010329J