Internet Electronic Journal of Molecular Design **2004**, *3*, 544–559 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

September 2004, Volume 3, Number 9, Pages 544–559

Editor: Ovidiu Ivanciuc

Proceedings of the Internet Electronic Conference of Molecular Design, IECMD 2003
November 23 – December 6, 2003
Part 3

Classification of Polar and Nonpolar Aquatic Pollutants Using Simple Descriptors. Differences between Polarity Prediction and Narcosis Classification

Guido Sello

Dipartimento di Chimica Organica e Industriale, Università degli Studi di Milano, via Venezian 21, 20133 Milano, Italy

Received: October 31, 2003; Revised: May 5, 2004; Accepted: May 21, 2004; Published: September 30, 2004

Citation of the article:
G. Sello, Classification of Polar and Nonpolar Aquatic Pollutants Using Simple Descriptors. Differences between Polarity Prediction and Narcosis Classification, *Internet Electron. J. Mol. Des.* **2004**, *3*, 544–559, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Classification of Polar and Nonpolar Aquatic Pollutants Using Simple Descriptors. Differences between Polarity Prediction and Narcosis Classification[#]

Guido Sello *

Dipartimento di Chimica Organica e Industriale, Università degli Studi di Milano, via Venezian 21, 20133 Milano, Italy

Received: October 31, 2003; Revised: May 5, 2004; Accepted: May 21, 2004; Published: September 30, 2004

Internet Electron. J. Mol. Des. **2004**, *3* (9), 544–559

Abstract

Motivation. The problem of toxicity prediction is mainly related to the necessity of processing many data that most of the time come from different sources and have different biological meaning. Often, the real mechanism of action of a toxicant is unclear or difficult to reproduce; in addition, a chemical compound exercises its toxic action through many steps that depend both on its structure and on the specific environment where it is acting. In this perspective, the classification of compounds can be of great help because decreases the number of the alternatives to those specific of that class, allowing a more focused analysis. The classification of narcotic pollutants into polar and nonpolar sets is certainly an important aspect of this type of problems.

Method. Object classification requires two principal components: the selection of the molecular descriptors and the choice of the classification algorithm. The calculation of the molecular descriptors is performed using our own approach that is based on empirical equations. We calculated three descriptors (Helc, HQ^+, Elcdif) that are used in pairs (Helc and Elcdif, or HQ^+ and Elcdif). Using two classification algorithms, a classical neural network and a tree neural network, we analyze two compound sets; the first contains 190 narcotic pollutants (114 nonpolar and 76 polar), the second contains 30 pollutants (20 nonpolar, 5 polar, 5 acetylcholinesterase inhibitors). In a broad sense, the first set is used as training set and the second as test set.

Results. The use of simple descriptors allows for a very good classification of narcotic pollutants demonstrating that it is not necessary to use high–level theories to make simple operations. On the contrary, much work is still required to obtain an acceptable theoretical prediction; part of it is definitely on the modelers' side, but the rest concerns a better rationalization of the experimental data without which any model will have problems.

Conclusions. Classification of narcotic pollutants into polar and nonpolar sets is required to ease the QSAR treatment of their toxic effects. However, there still remain many questions on the validation of theoretical models using only experimental data.

Keywords. Aquatic toxicity; narcotic pollutants; compound classification; empirical descriptors; experimental classification.

Abbreviations and notations

Elcdif, chemical potential difference between non hydrogen atoms	HQ^+, residual atomic charge on hydrogen atom
Helc, residual chemical potential on hydrogen atom	LUMO, Lowest unoccupied molecular orbital

[#] Presented in part at the Internet Electronic Conference of Molecular Design, IECMD 2003.

* Correspondence author; E–mail: guido.sello@unimi.it.

1 INTRODUCTION

The solution of the problems posed by environmental toxicity of chemicals requires many lines of reasoning; however, the first objective of any study in this field must be directed toward the determination of the toxicity level of the compounds. The first possibility is the experimental measure of the toxicity values, the second is their theoretical prediction. Aquatic toxicity is considered well–represented by the toxic effects of chemicals on fathead minnow [1–4], whose behaviour in the presence of chemical compounds has been studied and used to classify modes of action [2]. One of these last is the narcosis effect that is the consequence of the incorrect functioning of cell membranes. There are two different narcosis classes and three different behaviour syndromes; the class is assigned by joint toxic action bioassays with octanol (narcosis I or base–line narcosis) and phenol (narcosis II or polar narcosis); the syndromes are assigned by visual inspection and are: in the first mode the fish shows depressed locomotor activity with scarce response to outside stimuli (syndrome type I); in the second mode the fish is hyperactive and highly sensitive to outside stimuli (syndrome type II); in the third mode the fish shows a high incidence of convulsions, spasms, tetany, scoliosis, lordosis, and/or hemorraging in the vertebral column (syndrome type III). Two classes of compounds have been correlated to the two narcosis classes: non–polar and polar narcotic compounds, thus the classification of compounds in the correct class permits the more appropriate use of prediction models. However, no similar assignment to syndrome type has been done. Very recently, Ivanciuc [5] developed a highly efficient system for the classification of non–polar and polar narcotic pollutants, using two quantum descriptors (atomic charges on hydrogen atoms and the energies of the lowest unoccupied molecular orbital) and a new class of algorithms, Support Vector Machines (SVM) [6]. In this paper, we are going to use simpler descriptors and two diverse clustering methodologies: Classification Neural Network and Classification Tree [7]. In addition, the second method allows for the sub classification of objects giving further insights into the compound relation. Finally, we will use a second compound set to discuss the differences that are still present between theoretical and experimental models.

2 MATERIALS AND METHODS

2.1 Chemical Data

2.1.1 Descriptors calculation. Charges and residual chemical potentials

The choice of descriptors is the critical point when developing a model. Often, we have too many potential descriptors whose selection will affect the outcome of the model. Where it is possible to make a hypothesis on the mechanism of the biological action we can support our choice on that ground. However, in all other cases the choice is guided by our mere judgment. In the present case, we accept the choice made by other authors [5,8–9] that select charge and molecular

nucleophilicity (represented by LUMO energy) as good descriptors of molecule "polarity". It is clear that this model is quite simple, but the literature results in the compound classification are impressive. Taken into consideration the model simplicity we would like to test if the use of the same descriptors calculated at lower theory level works similarly.

We are going to use our own program [10–12] for atomic descriptor calculation to obtain: (*a*) the highest positive charge on a hydrogen atom, as used by Ivanciuc [5]; or, the highest residual chemical potential on a hydrogen atom, representing the same effect; (*b*) the highest difference in residual chemical potentials between non hydrogen atoms, as a substitute of LUMO energy, the descriptor of the molecule nucleophilicity.

The method used to perform calculation is based on an original approach that uses chemical potential equalization as driving theory. At the end of the calculation each atom has two descriptors of its electronic state: the partial atomic charge and the residual chemical potential. These two data are correlated, but they have different resolution; atomic charge is defined at the 10^{-3} electron level, whilst the residual potential is defined at the 10^{-2} electric potential level. Often, it happens that two atoms with the same residual potential have different atomic charge; this is particularly visible for hydrogen atoms. The choice of the hydrogen atom and of the heavy atom pair to consider is straightforward: the hydrogen atom with the highest positive charge and the pair of connected atom with the highest difference in chemical potential are selected.

2.2 Biological Data

We are going to use two sets of biological data, both concerning narcosis effects. The first set is exactly the same set used by Ivanciuc [5]; it will be the training set of the analysis and its use should allow for a comparison to the Ivanciuc's result. The second set is a smaller set (30 compounds) selected from the list of Russom *et al*. [2] showing different experimental toxic effects; it will be used to test the classification model and to discuss the differences between experimental and calculated toxicities.

2.3 Classification Algorithms

Object classification can be achieved using many different models. We chose two different approaches that are representative of two different techniques [7]. The first is a classical artificial neural network that partitions a set of objects into the assigned classes and validates the results; the second is a hierarchical method that grows a tree where each final leaf contains a subset resulting from successive splitting operations. The two models have been applied to two descriptor sets (HQ^+/Elcdif; Helc/Elcdif), to two compound sets (training set (190), test set (30)). In all cases a further run has been performed using randomized class assignment in order to check the algorithms predictivity; the results show that the randomized sets do not give reliable classifications.

2.3.1 Classification NN

It is a very basic implementation of FeedForward – BackPropagation Neural Network, used for prediction and classification problems; the corresponding algorithm is freely available [7]. The network was tested using different combinations of the input parameters: input and output nodes fixed at 2 and 1, respectively; hidden layers 1; layer sizes 2, 3, 4; learning parameter between 0.1 and 0.9; momentum between 0.01 and 0.1; epochs between 200 – 1000; validation set containing 10 or 20 % of total cases, randomly selected; training mode chosen as "sequential". The best results were obtained with the following settings.

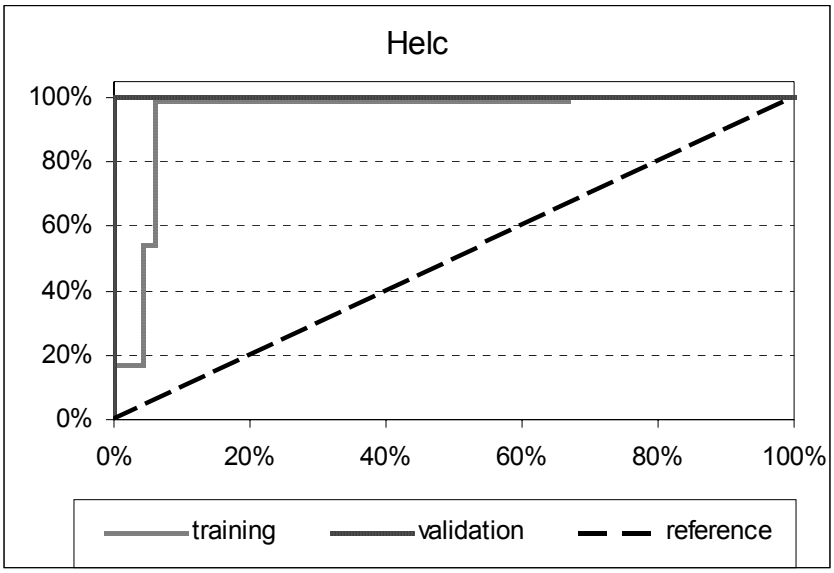


Figure 1. ROC curve obtained with the Helc descriptor.

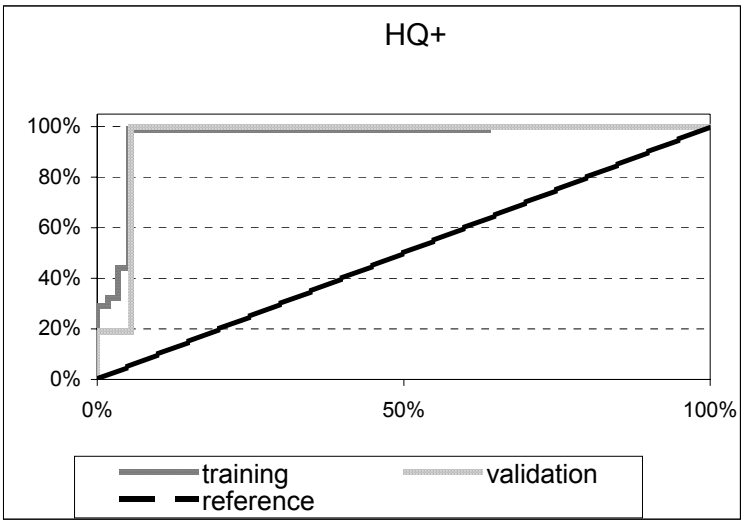


Figure 2. ROC curve obtained with the HQ^+ descriptor.

Neural network classification using Helc: inputs number = 2, hidden layers = 1, layer size = 3, learning parameter = 0.9, momentum = 0.1, cycles = 500, training mode = sequential, validation =

20%–random, ROC (Receiver's Operating Curve) is presented in Figure 1. Neural network classification using HQ^+: inputs number =2, hidden layers = 1, layer size = 3, learning parameter = 0.7, momentum = 0.1, cycles = 500, training mode = sequential, validation = 20%–random, ROC is presented in Figure 2. In all cases a randomized input, i.e. randomly assigning cases to classes, has been used to check chance classification; results show the expected complete unpredictability.

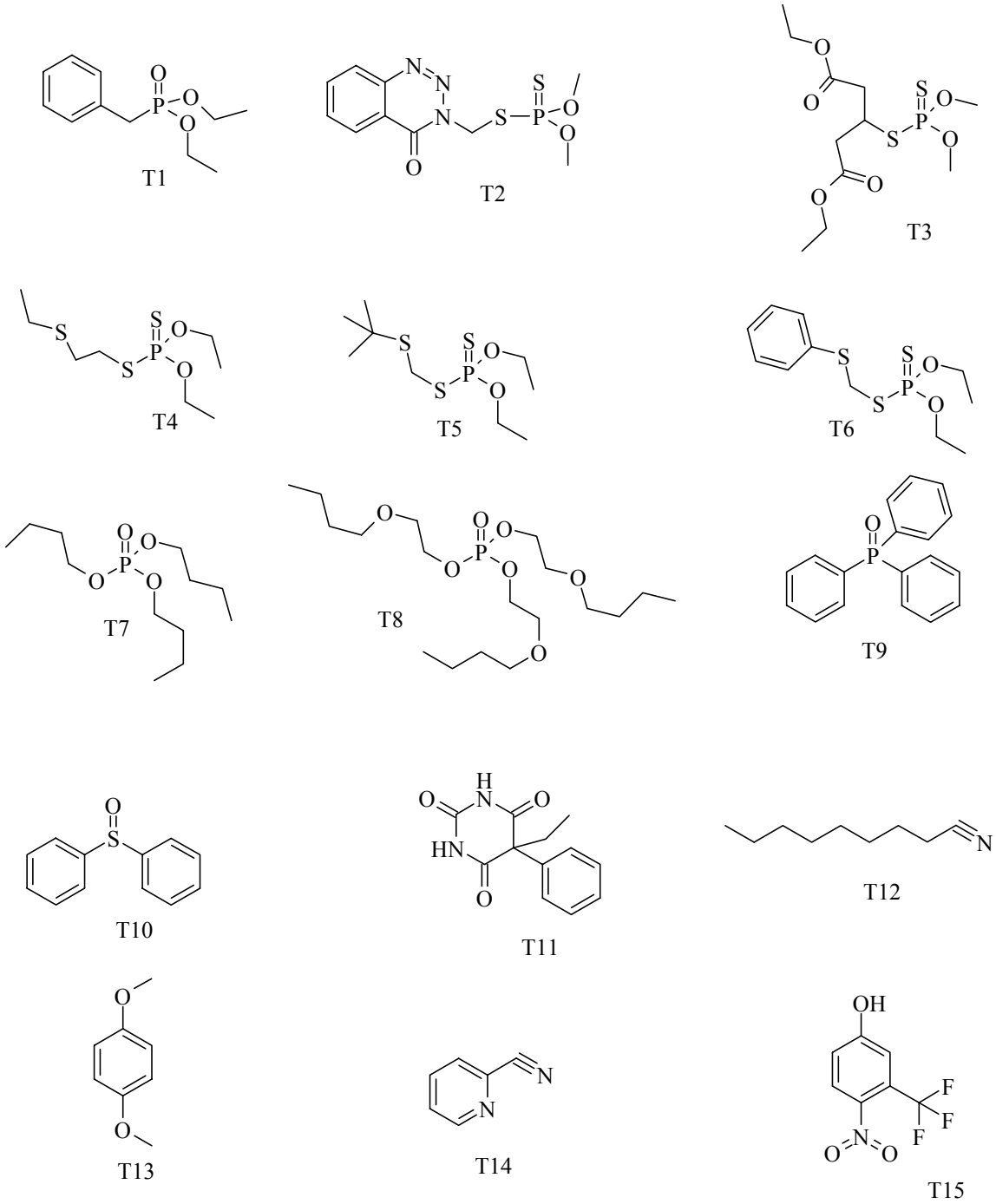


Figure 3. Compounds in the test set.

It must be pointed that many other settings give results of the same quality, but the cited settings

are efficient and reliable. The networks arrived at convergence before reaching the maximum number of epochs (310 and 60 respectively). The wrong predictions are always the same when the wrong–predicted case is part of both the training and the validation set. This means that the wrong predictions depend on the variable values and not on the algorithm performance.

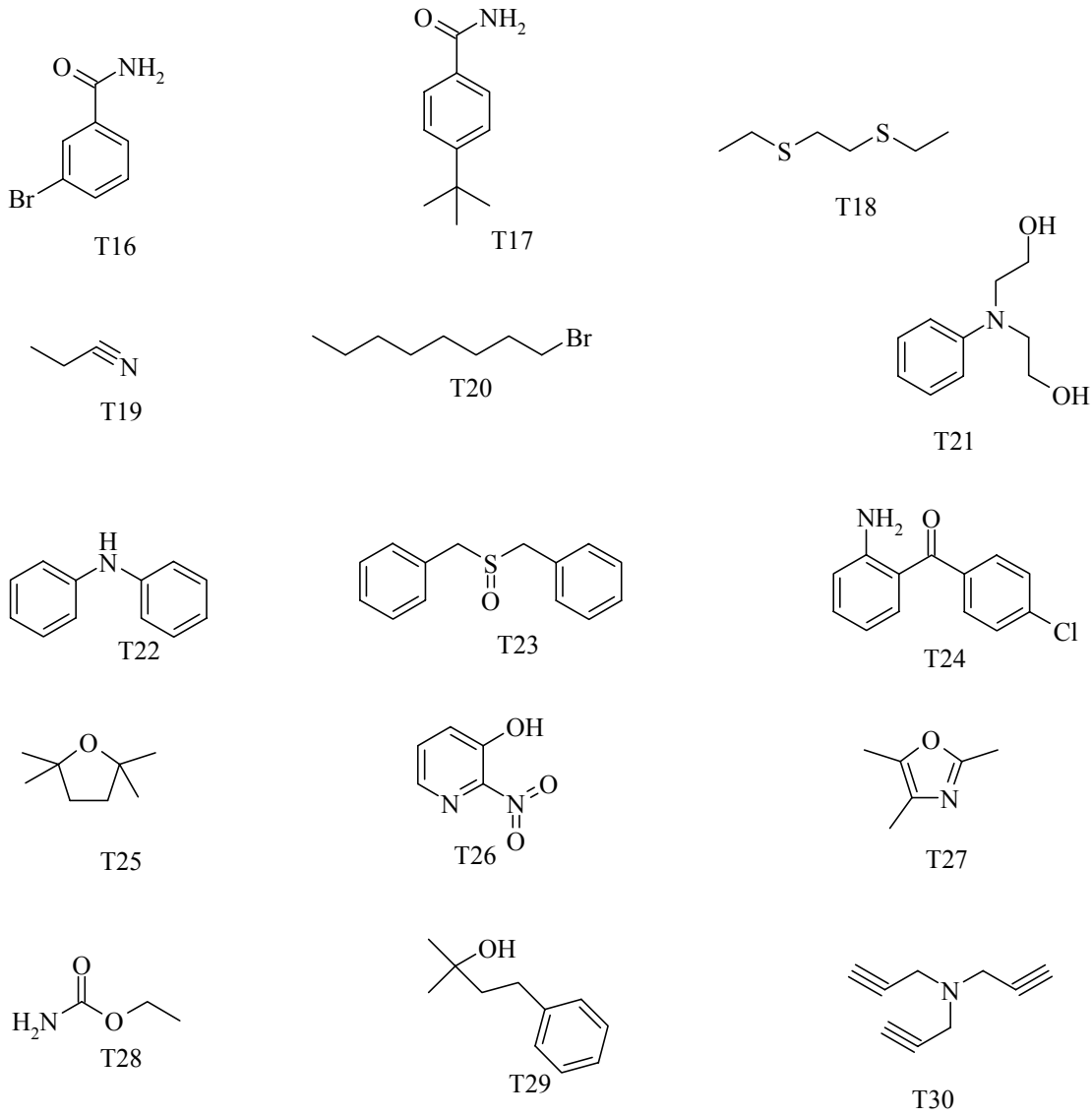


Figure 4. Compounds in the test set.

2.3.2 Classification tree

It is a classification model that (*a*) uses C4.5 (see Appendix 2) algorithm by Ross Quinlan [13]; (*b*) has a Node Splitting Criterion that uses Entropy based criterion to select the split. While growing the tree, at any point a predictor is chosen to split a node such that the Information Gain is maximized after the split. As specified in C4.5, it presently uses the *Gain Ratio* (*Gain Ratio* = *Gain*/ *Split Info*) to choose the split. (*c*) has a Stopping Criteria that stops splitting a node and declares it as a leaf node if any one of the following criterion is met: (1) number of records in the

node is less than some pre–specified limit; (2) purity of the node is more than some pre–specified limit *p*. This means that the proportion of records in the node with class equal to the majority class is *p* or more. (3) depth of the node is more than some pre–specified limit. (4) predictor values for all records are identical. (*d*) has a Tree Pruning based on the pessimistic error rate at the node. If the pessimistic error rate of a node is less than that of the subtree rooted at that node, the node is pruned. If we fail to prune a node – none of its predecessors is pruned. (*e*) has a Rule Generation according to the methods mentioned in C4.5. The corresponding algorithm is also freely available [7]. The adopted criterion for stopping the split of trees was 20% or less nodes minimum. The generated trees are reported in Figure 5 and 6.

3 RESULTS AND DISCUSSION

3.1 Polarity Prediction

The training set is exactly that used by Ivanciuc [5], thus we are not explicitly reporting its components. The compounds in the test set are shown in Figures 3 and 4. They have been selected from Russom *et al*. [2], 23 molecules are reported to have narcosis I effect, 2 narcosis II effect (T14, T15), 5 are acetylcholinesterase inhibitors (T2–T6).

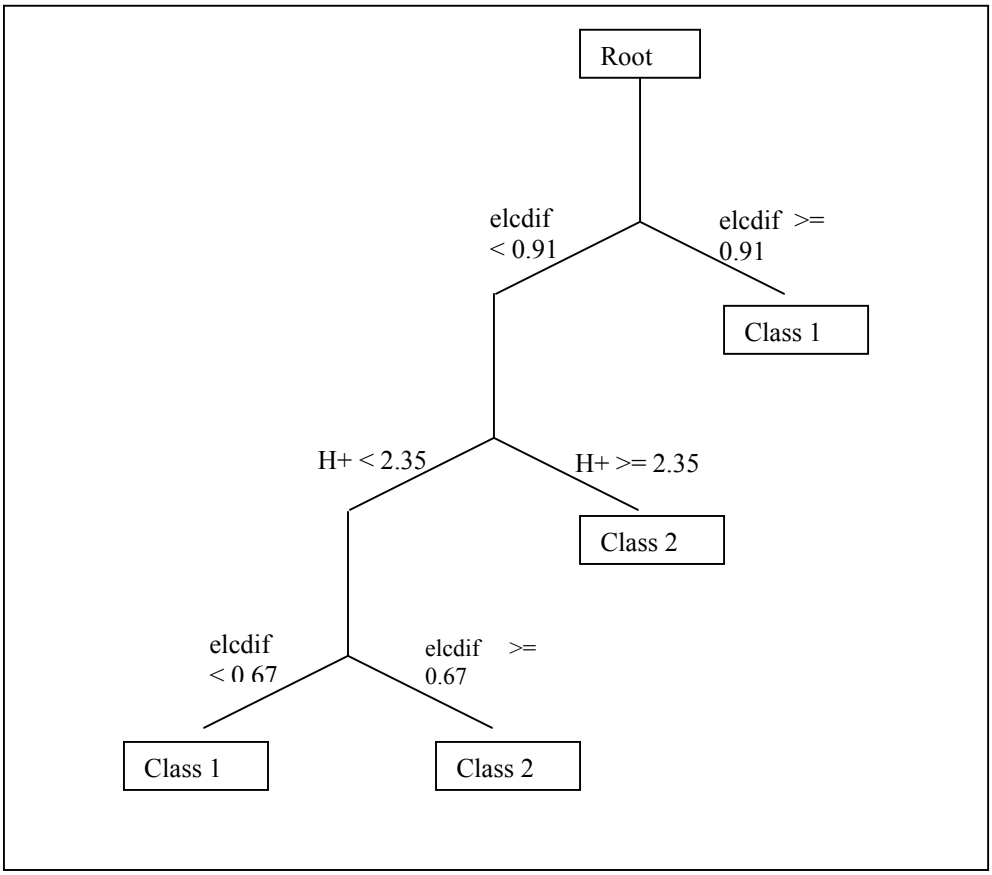


Figure 5. Tree generated using Helc descriptor.

For all the 220 molecules we have calculated the atomic descriptors and selected those useful for the model, as reported in 2.1.1. In particular, we selected the greatest chemical potential difference between connected non–hydrogen atoms, the greatest chemical potential on a hydrogen atom, and the greatest residual charge on a hydrogen atom. The first descriptor is greater if the two connected atoms are electronically different; it can be interpreted as the force acting on an external positive charge, i.e. it represents the Q^+ accepting power. The second descriptor has the same meaning concerning Q^- accepting power by hydrogen atoms; it is thus related to hydrogen bond formation. The last descriptor is the same used by Ivanciuc [5], but calculated by our method. The reason behind the alternative use of the second and third descriptors is strictly related to the calculation method. In fact, depending on the molecular neighborhood the residual charge can be different on hydrogens that have the same residual chemical potential; thus, if we consider that the hydrogen bond power is only an electrostatic effect the third descriptor is the right one, but if the hydrogen bonding involves an electron movement the second descriptor must be used.

3.2 Narcosis Classification

This biological effect is experimentally measured by joint toxic action measurements after the treatment by the chemical at different concentrations and in different combinations. There is ample literature on this matter and we are not going to discuss the different data or protocols. Nevertheless, it must be emphasized that there exists some discrepancies between single laboratory results and their interpretation. This is important because the discussion on the models must consider the variability of the biological data.

In principle, there are several narcosis classes that can be roughly divided into base–line narcosis, polar narcosis, and ester narcosis, this last can be merged with the second type. In addition, there are several confidence levels in the class allocation. We generally accept the Ivanciuc's interpretation when studying the training set, whereas we are going to discuss the results of the test set considering the Russom *et al*. indications.

3.2.1 Training set and test set

In the following two Tables the results of the models are reported.

Classical NN classification gives a result that is in overall agreement with Ivanciuc's [5]. There are 5 misclassified objects using both Helc and HQ^+ (54, 164, 181, 182, 183) with respect to 11 objects in Ivanciuc (21, 23, 32, 47, 60, 62, 68, 69, 156, 157, 164). It is worth to note that, excluding compound 164, the misclassified objects are different. In our case, misclassification is related to Elcdif in all cases but for compound 164 (here the HQ^+, or Helc, is responsible). The analyses have been performed at least five times randomly selecting a 10% or 20% validation set and the result has always been the same. The test set has been classified using the obtained models and the result showed slightly better for the HQ^+ descriptor. Not all the test compounds are in the expected class,

but we have 8 or 5 misclassifications (T10, T12, T15, T19, T22, T26, T27, T30), or (T10, T15, T22, T26, T30). It is remarkable that many misclassifications present in HQ^+ are also present in Helc.

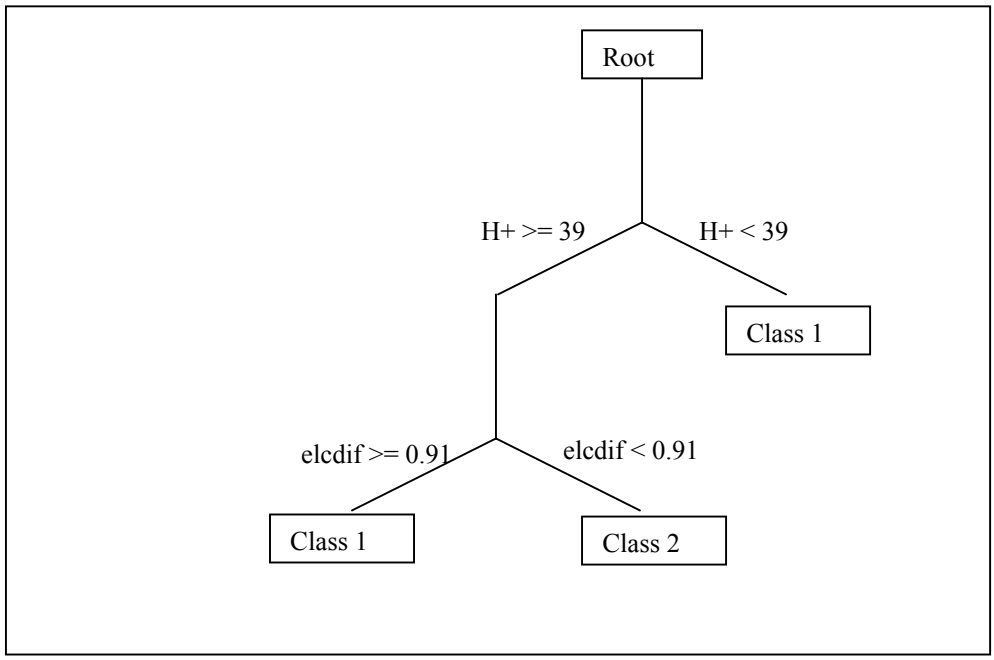


Figure 6. Tree generated using HQ^+ descriptor.

Table 1. Classical–NN representative results

		Training [a]	[b]	Misclassified [c]	[d]	Validation [e]	20% [f]	Misclassified [g]	20% [h]
Helc+Elcdif	class 1	101	84	1	1	12	29	0	0
	class 2	64	62	4	4	8	10	0	0
Test [i]	class 1	21		7					
	class 2	1		1					
Rand [j]	class 1	0		84		0		10	
	class 2	86		0		10		0	
HQ+Elcdif	class 1	96	92	1	1	17	21	0	0
	class 2	69	55	4	3	3	17	0	1
Test [i]	class 1	24		4					
	class 2	1		1					
Rand [j]	class 1	87		0		7		0	
	class 2	0		83		0		13	

[a] Number of cases correctly classified in the training set, with 10% of the cases in the validation set. [b] Number of cases correctly classified in the training set, with 20% of the cases in the validation set. [c] Number of cases misclassified in the training set, with 10% of the cases in the validation set. [d] Number of cases misclassified in the training set, with 20% of the cases in the validation set. [e] Number of cases correctly classified in the validation set, containing 10% of the cases. [f] Number of cases correctly classified in the validation set, containing 20% of the cases. [g] Number of cases misclassified in the validation set, containing 10% of the cases. [h] Number of cases misclassified in the validation set, containing 20% of the cases. [i] Results for the test set (compounds T1–T30). [j] Results for the training set with random assignment of cases to the classes.

Tree classification gives a result that is very similar to the previous one. Both Helc and HQ^+ show a small number of misclassifications (5 and 6, respectively) of the same compounds (54, 164, 181, 182, 183, *75*) in very good agreement with the Classical NN. This demonstrates that the two

classification methods have very similar behaviour, as expected. This fact is confirmed by the test set that gives similar misclassifications (T10, T12, T14, T15, T19, T22, T26, T27, T30). In this case the class values have been calculated using the rules that the approach produces during the training. These rules are, in order of application: (1) Elcdif < 0.91; (2) Helc < 2.35; (3) Elcdif < 0.67; and (1) HQ^+ < 39; (2) Elcdif< 0.91.

Table 2. Tree classification results

		Training [a]	Misclassified [b]	Nodes [c]	Levels [d]
Helc+Elcdif	class 1	110	4	4	4
	class 2	75	1		
Test [e]	class 1	22	6		
	class 2	1	1		
Rand [f]	class 1	69	45	10	11
	class 2	51	25		
HQ+Elcdif	class 1	110	4	3	3
	class 2	74	2		
Test [e]	class 1	24	4		
	class 2	0	2		
Rand [f]	class 1	70	44	10	8
	class 2	46	30		

[a] Number of cases correctly classified in the training set. [b] Number of cases misclassified in the training set. [c] Number of nodes in the tree. [d] Number of levels in the tree. [e] Results for the test set (compounds T1–T30). [f] Results for the training set with random assignment of cases to the classes.

Table 3. Statistical results of compound classification using NN and Tree algorithms

	TP [a]	FN [a]	TN [a]	FP [a]	Accuracy [a]	TP [b]	FN [b]	TN [b]	FP [b]	Accuracy [b]
Helc+Elcdif [c]	113	1	72	4	0.97	113	1	72	4	0.97
HQ+Elcdif [c]	113	1	72	4	0.97	113	1	72	4	0.97
Helc+Elcdif [d]	110	4	75	1	0.97					
HQ+Elcdif [d]	110	4	74	2	0.97					

[a] True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP) for L10%O analyses. [b] True Positive (TP), False Negative (FN), True Negative (TN), and False Positive (FP) for L20%O analyses. [c] Neural network classification results. [d] Tree classification results

Table 4. Chemical equivalent classes of the training set compounds

Using Helc			
Cases	*Misclassified*	*Chemical equivalent*	*Misclassified chemical equivalent*
1–60	*54*	alcohols, ketones, esters, ethers	*diphenyl ether*
61–114		halides, hydrocarbons	
115–130; 156–157		nitro compounds, pyridine and quinoline	
131–190	*164* and *181–183*	phenols, anilines	*N,N–dimethyl aniline* and *fluoro anilines*
Using HQ^+			
Cases	*Misclassified*	*Chemical equivalent*	*Misclassified chemical equivalent*
1–32, 56, 60		alcohols, furan, 2–hydroxy–4–methoxy acetophenone	
33–114	*54* and *75*	ketones, esters, ethers, halides, hydrocarbons	*diphenyl ether* and *trichloroethene*
115–190	*164* and *181–183*	nitro compounds; pyridine and quinoline, phenols, anilines	*N,N–dimethyl aniline* and *fluoro anilines*

3.2.2 Subclassification

Compound classification in non–polar and polar narcotics is definitely interesting because it should allow for the use of the appropriate QSAR. However, due to the extended diversity of compounds it could be also interesting to divide them in more classes with the objective of a better prediction. This can be done using the Tree clustering method. In the Helc case we have four terminal leaves, whereas in the HQ^+ we have only three terminal leaves; as a consequence we obtain four or three compound subsets. They are sketched in Table 4.

In Table 5 we have reported the misclassified compounds of the test set. Depending on the descriptors and on the considered class (main class or syndrome class) we have from 5 to 11 misclassifications. In addition, even some of the constantly misclassified compounds (T10, T14, T15, T22, T26) i.e. T15, T22, and T26, have a hydrogen atom sufficiently polar to classify them in the right polarity class but in the wrong narcosis class. Thus, taking into consideration all these facts our method can be considered acceptable also in the classification of the test compounds.

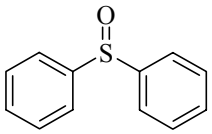
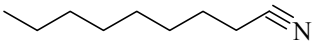
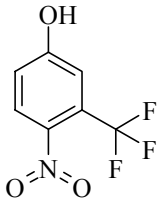
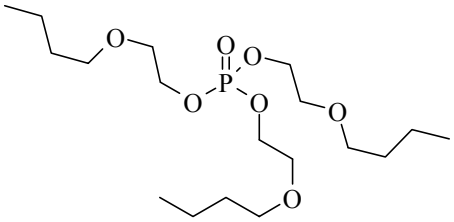
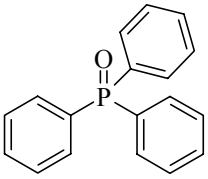
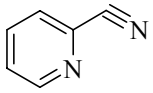
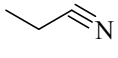
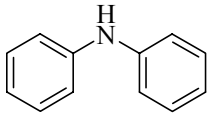
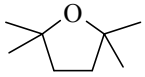
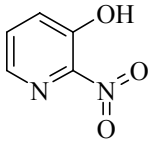
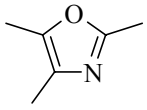
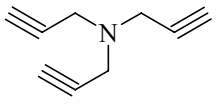
In order to compare our result to that by Ivanciuc's method [5] it is interesting to note the variable ranges that are:

Present work			Ivanciuc's data
HQ^+	0 – 244	HQ^+	0 – 397
Elcdif	0 – 1.74	E_{LUMO}	–1.49 – 3.78
Helc	0 – 2.57		

It is clear that using quantomechanical approaches the variability of the values is higher; however, this variability has an influence only in the case of E_{LUMO}, because the HQ^+ values are very similar. Nevertheless, the sensitivity of the molecular orbital methods to small variations in the molecular geometry is well known and, thus, the meaning of small variations of the E_{LUMO} values are unimportant.

The consequence is that it is seldom possible to predict and to understand the misclassifications; for example, in the case of 3–furanmethanol the E_{LUMO} value is the cause of the wrong prediction when the power of hydrogen bond forming is probably due to the methanol part, only. Our values, on the contrary, allow for an immediate understanding of the misclassifications; for example, in the diphenyl ether case the Elcdif is the cause of the wrong prediction and it is related to the absence of sufficiently different atomic chemical potential (here the C–O bond is less polar than in alkyl compounds).

Table 5. Misclassified or partially misclassified compounds in the test set

Compound	MOA class	Behaviour syndrome	Predicted class[a]
T10 	Narcosis I	Type I	Polar
T12 	Narcosis I	Type I	Polar/Non–polar
T15 	Narcosis II	Type II	Non–polar
T8 	Narcosis I	Type II	Non–polar
T9 	Narcosis I	Type II	Non–polar
T14 	Narcosis II	Type III	Polar
T19 	Narcosis I	Type n.d.	Polar/ Non–polar
T22 	Narcosis I	Type n.d.	Polar
T25 	Narcosis I	Type II	Non–polar
T26 	Narcosis I	Type I	Polar
T27 	Narcosis I	Type I	Polar/ Non–polar
T30 	Narcosis I	Type I	Polar

[a] When two classes are reported the assignment depends on either the descriptor or the method.

Many other authors have used the same set, or even more extended sets, and prepared models for both classifications [9,14,15] and quantitative [8,16] correlations. These last are not directly comparable to the present work; however, some of the predictions are in good agreement with experimental data. Among classification approaches we must cite Ren and Schultz studies [9,14]; they classify compounds using logP, HOMO, LUMO, and soft electrophilicity, descriptors in different combinations. The obtained results are comparable to Ivanciuc's and to the present ones. The main difference is the use of the water-octanol partition coefficient that intends to describe the transport mechanism. This last is a well-known important experimental characteristic affecting the activity of chemicals; nevertheless, it seems that the polarity description of compounds is sufficient for classification purpose.

3.2.3 Experimental data and theoretical predictions

The final part of this paper will be concerned with the differences between theoretical predictions and experimental fish behavior classifications. We can evidence two different classifications in experimental data: main classification that is related to the mode of action, secondary classification that is related to fish behavior (syndrome type). Thus, we have compounds that, even if members of the same main class, show different behavioral syndromes. In addition, the confidence level of the class assignments can greatly differ making quite difficult the interpretation of the data (four levels of confidence). In contrast, in the present classification model we are only separating non–polar and polar compounds in the assumption that these two classes are representative of narcosis I and II MOAs, respectively. In addition, in the test set we voluntarily introduced compounds (T2–T6) that cannot be properly classified as narcotics. The comparison between experimental and calculated data is therefore complicated.

First, we are going to comment on the 190 compounds of the training set. Running through the table presented by Russom *et al*. [2] we can easily identify some compounds that are classified in behavioural classes different from those predicted by calculation. For example: (*a*) trichloroethene causes a narcosis I class II syndrome, classified as non–polar in ref. 5; (*b*) *N*,*N*–dimethylaniline causes a narcosis I class I syndrome, classified as polar in ref. 5; (*c*) 4–ethylaniline causes a narcosis I class I syndrome, classified as polar in ref. 5; (*d*) 4–chlorophenol, 4–methoxyphenol, and pyridine, cause a narcosis II class III syndrome, classified as polar in Ref. [5].

It is evident that the fish reaction to chemicals is more complex than that predicted by models. Nevertheless, the majority of the compounds are inserted in the correct MOA class. In our approach, the misclassified compounds are: trichloroethene, diphenylether, *N*,*N*–dimethyl aniline, α,α,α,4–tetrafluoro–3–methylaniline, α,α,α,4–tetrafluoro–2–methylaniline, pentafluoroaniline. *N*,*N*–dimethylaniline is, however, classified as non–polar narcotic in Ref. [2], thus its misclassification cannot be regarded as a true error. In contrast, the other compounds contain special atoms (fluorine and chlorine) that are strongly polar and that in particular position and amount affect the compound total polarity. Nevertheless, they represent a true limit of the model.

The test set is even more complicated. We find ~11/30 misclassifications (in the worst case scenario), but we must take into consideration that here the class assignment is done following Russom *et al*. [2] In fact considering only main class membership the number of misclassifications decreases to 5 cases. In contrast, the classification of compounds T2–T6 demonstrates that the models cannot discriminate between different modes of action. In other words, we still need some experimental pre–elaboration (*e.g*. the assignment of the MOA) to be confident in the results. The insertion of compounds T2–T6 into the non–polar class can be interesting, but we must be careful in affirming that a non–polar fish toxicant automatically follows the narcosis I MOA. We have difficulties in estimating the relative value of experimental fish behavior and of calculated prediction. The experimental results have different levels of confidence, as mentioned, but they are real effects. The experimental assignment of the compounds to the different MOAs and syndrome types is performed using well–defined experimental protocols that use many different tests, from LC50 ratio to joint toxic action, to changes in behavior and morphology. On the other hand, calculated predictions only use a limited number of molecular descriptors that cannot describe all the experimental data but can only classify compounds by their polarity (or better by the polarity of single atoms or bonds). They are self consistent and have the same reliability, but they can represent an underestimation of the reality. Compound classification is only a first step toward the quantitative assessment of toxicity; we have therefore the chance of getting further corrections in the successive analysis.

A similar conclusion can be reached for acetylcholinesterase inhibitors; in the models they are classified in class 1 (nonpolar toxicants). However, their mode of action is completely different and should follow a different classification scheme. This fact is fundamental because it indicates that the

polarity of a compound is not the only factor to consider when predicting its MOA and toxicity.

4 CONCLUSIONS

Classification of compounds can represent a highly effective way to separate chemicals into specific sets that can then be analyzed by specific models. The choice of the descriptors useful to perform the classification is a critical point that requires attention both on the correspondence with physical properties and on the needed theory level. The choice of the classification algorithms is less crucial if the method is robust enough; a special attention can be dedicated to the selection of clustering algorithms with the aim of automatically sub classifying the compounds. Finally, the numerous details of the experimental data must be accurately considered to prevent invalid evaluation of the model performance.

Appendix 1

Rule Generation in C4.5

Unlike many algorithms that employ a separate-and-conquer approach to selecting a rule set that explains the training cases, C4.5 extracts rules from existing decision trees.

The objective of the procedure is to end with an ordered set of if -then rules of the form: if LHS then RHS where LHS is a combination of attribute-value conditions and RHS is a class assignment. During classification of an instance, the first rule that meets with all instance conditions specified in the LHS “fires”", and the class label specified in the RHS is assigned to the instance.

Because the hypothesis space is usually not completely included in the rule set, a default class is also present — this is the class label assigned to any instances that are not covered by any rule in the rule set. The four steps in the process of rule generation are:

(1) Pruning. First, an initial rule set is built by translating each path from the tree root to a leaf into a rule, where each condition in the LHS of the rule corresponds to an internal node of a path in the decision tree. All rules are examined, and conditions are removed from the LHS if they do not appear to contribute to the accuracy of the rule, giving a more general, “pruned”" rule that is added to the intermediate rule set if it is not already present.

(2) Selection. The pruned rules built in the previous step are grouped into class rule sets, one for each of the k classes in the training set (i.e., the RHS of each rule in subset (i) is class C_i). Each of these rule sets is examined again to select a subset of the rules that maximize the predictive accuracy.

(3) Ordering. The k rule sets are ordered according to the frequency of false positive errors and a default class is chosen by assigning the most frequent class of uncovered cases.

(4) Evaluation. The rule set is evaluated against the training set to determine any rules increasing classification error. The “worst” such rule is removed and evaluation is repeated until no further improvement is possible.

5 REFERENCES

[1] S. Karabunarliev, O. G. Mekenyan, W. Karcher, C. L. Russom, and S. P. Bradbury, Quantum–Chemical Descriptors for Estimating the Acute Toxicity of Electrophiles to the Fathead Minnow (*Pimephales promelas*): An Analysis Based on Molecular Mechanisms, *Quant. Struct.–Act. Relat*. **1996**, *15*, 302–310.

[2] C. L. Russom, S. P. Bradbury, S. J. Broderium, D. E. Hammermeister, and R. A. Drummond, Predicting Modes of Toxic Action From Chemical Structure: Acute Toxicity in the Fathead Minnow (*Pimephales promelas*), *Environ. Toxicol. Chem*. **1997**, *16*, 948–967.

[3] A. P. Bearden and T. W. Schultz, Structure–Activity Relationships for *Pimephales* and *Tetrahymena*: A Mechanism of Action Approach, *Environ. Toxicol. Chem*. **1997**, *16*, 1311–1317.

[4] A. P. Bearden and T. W. Schultz, Comparison of *Tetrahymena* and *Pimephales* Toxicity Based on Mechanism of Action, *SAR QSAR Environ. Res*. **1998**, *9*, 127–153.

[5] O. Ivanciuc, Aquatic Toxicity Prediction for Polar and Nonpolar Narcotic Pollutants with Support Vector Machines, *Internet Electron. J. Mol. Des.* **2003**, *2*, 195–208, http://www.biochempress.com.

[6] A. Ben–Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, Support Vector Clustering, *J. Machine Learning Res.* **2001**, *2*, 125–137.

[7] The software used is freely available at http://www.geocities.com/adotsaha/index.html.

[8] E. Urrestarazu Ramos, W. H. J. Vaes, H. J. M. Verhaar, and J. L. M. Hermens, Quantitative Structure–Activity Relationships for the Aquatic Toxicity of Polar and Nonpolar Narcotic Pollutants, *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 845–852.

[9] S. Ren, Classifying Class I and Class II Compounds by Hydrophobicity and Hydrogen Bonding Descriptors, *Environ. Toxicol*. **2002**, *17*, 415–423.

[10] L. Baumer, G. Sala, G. Sello, Residual Charges on Atoms in Organic Structures: A New Algorithm for Their Calculation. *Tetrahedron. Comput. Method.,* **1989**, *2*, 37–46.

[11] L. Baumer, G. Sala, G. Sello, Residual Charges on Atoms in Organic Structures: A New Method for the Identification of Conjugated Systems and the Evaluation of Atomic Charge Distribution on Them. *Tetrahedron. Comput. Method.* **1989**, *2*, 93–103.

[12] L. Baumer, G. Sala, G. Sello, Residual Charges on Atoms in Organic Structures: Molecules Containing Charged and Backdonating Atoms. *Tetrahedron. Comput. Method.* **1989**, *2*, 105–118.

[13] J. R. Quinlan, C4.5: Program for Machine Learning, San Mateo: Morgan Kaufmann **1993**.

[14] S. Ren, T. W. Schultz, Identifying the mechanism of aquatic toxicity of selected compounds by hydrophobicity and electrophilicity descriptors. *Toxicol*. *Lett*. **2002**, *129*, 151–160.

[15] S. C. Basak, G. D. Grunwald, G. E. Host, G. J. Niemi, S. P. Bradbury, A comparative study of molecular similarity, statistical, and neural methods for predicting toxic modes of action. *Environ. Toxicol. Chem*. **1998**, *17*, 1056–1064.

[16] K. L. E. Kaiser, S. P. Niculescu, Using probabilistic neural network to model the toxicity of chemicals to the Fathead Minnow (*Pimephales Promelas*): a study based on 865 compounds. *Chemosphere* **1999**, *38*, 3237–3245.