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

Guido Sello*

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

Internet Electronic Conference of Molecular Design **2003**, November 23 – December 6

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 remains 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

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 effects: in the first mode the fish

shows depressed locomotor activity with scarce response to outside stimuli (narcosis I or base-line narcosis); in the second mode the fish is hyperactive and highly sensitive to outside stimuli (narcosis II or polar narcosis). Two classes of compounds have been correlated to the two modes: non-polar and polar narcotic compounds, thus the classification of compounds in the correct class permits the more appropriate use of prediction models. 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 still are present between theoretical and experimental models.

2 MATERIALS AND METHODS

2.1 Chemical Data

2.1.1 Calculation of descriptors. 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 opinion. 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 reported 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:

- 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
- the highest difference in residual chemical potentials between non hydrogen atoms, as a substitute of LUMO energy, the descriptor of the molecule nucleophilicity

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 is 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.

2.3.2 Classification tree

It is a classification model that

- uses C4.5 algorithm by Ross Quinlan. [13]
- 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 actually uses the *Gain Ratio (= Gain / Split Info)* to choose the split.
- 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.
  - Number of records in the node is less than some pre-specified limit.
  - 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.

- Depth of the node is more than some pre-specified limit.
- Predictor values for all records are identical.

o 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.

o has a Rule Generation according to the methods mentioned in C4.5.

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 Figure 1. They have been selected from Russom et al. [2], 20 molecules are reported to have narcosis I effect, 4 narcosis II effect (T8, T9, T15, T25), 1 narcosis III effect (T14), 5 are acetylcholinesterase inhibitors (T2-T6).

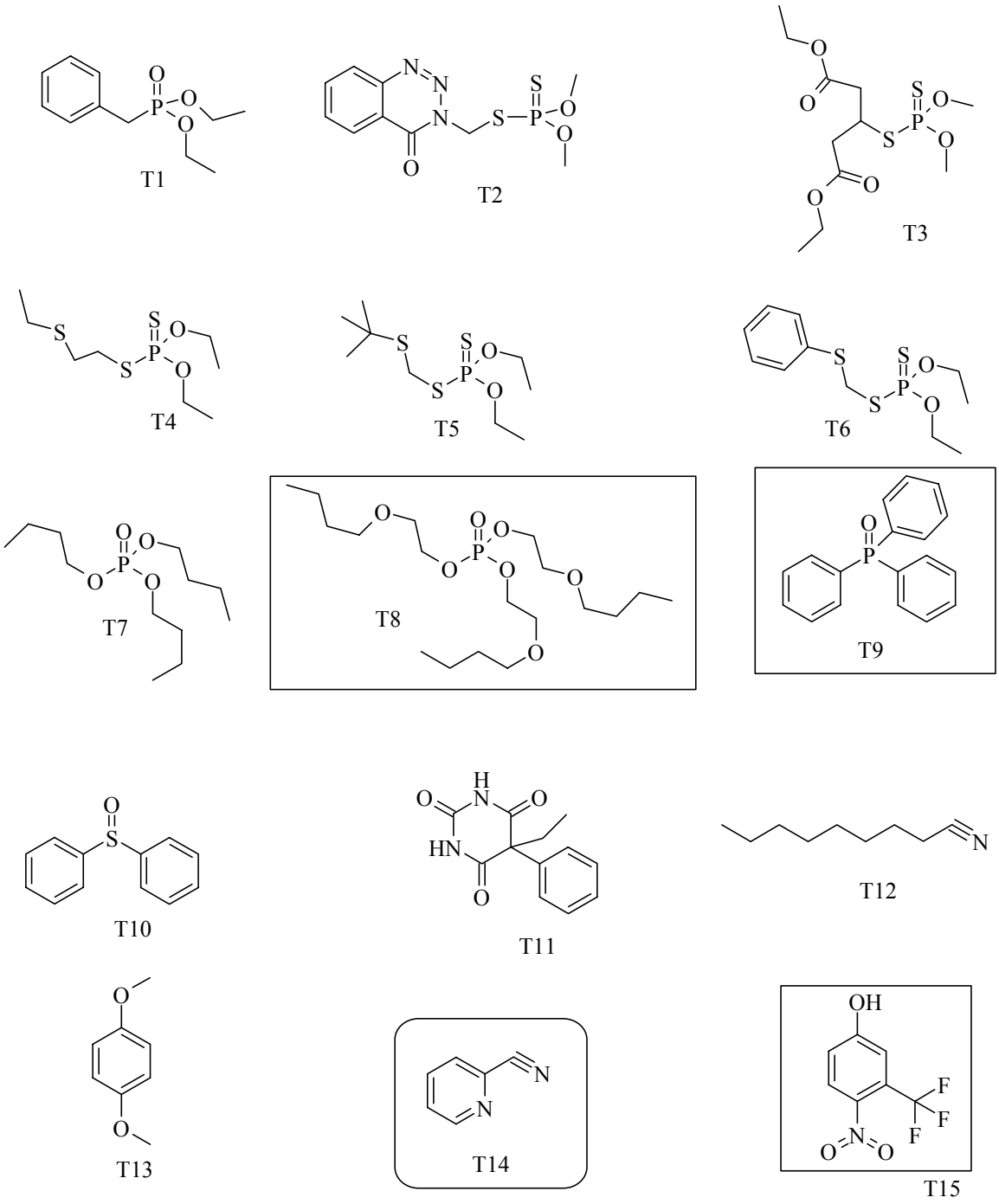

Figure 1. Compounds in the test set

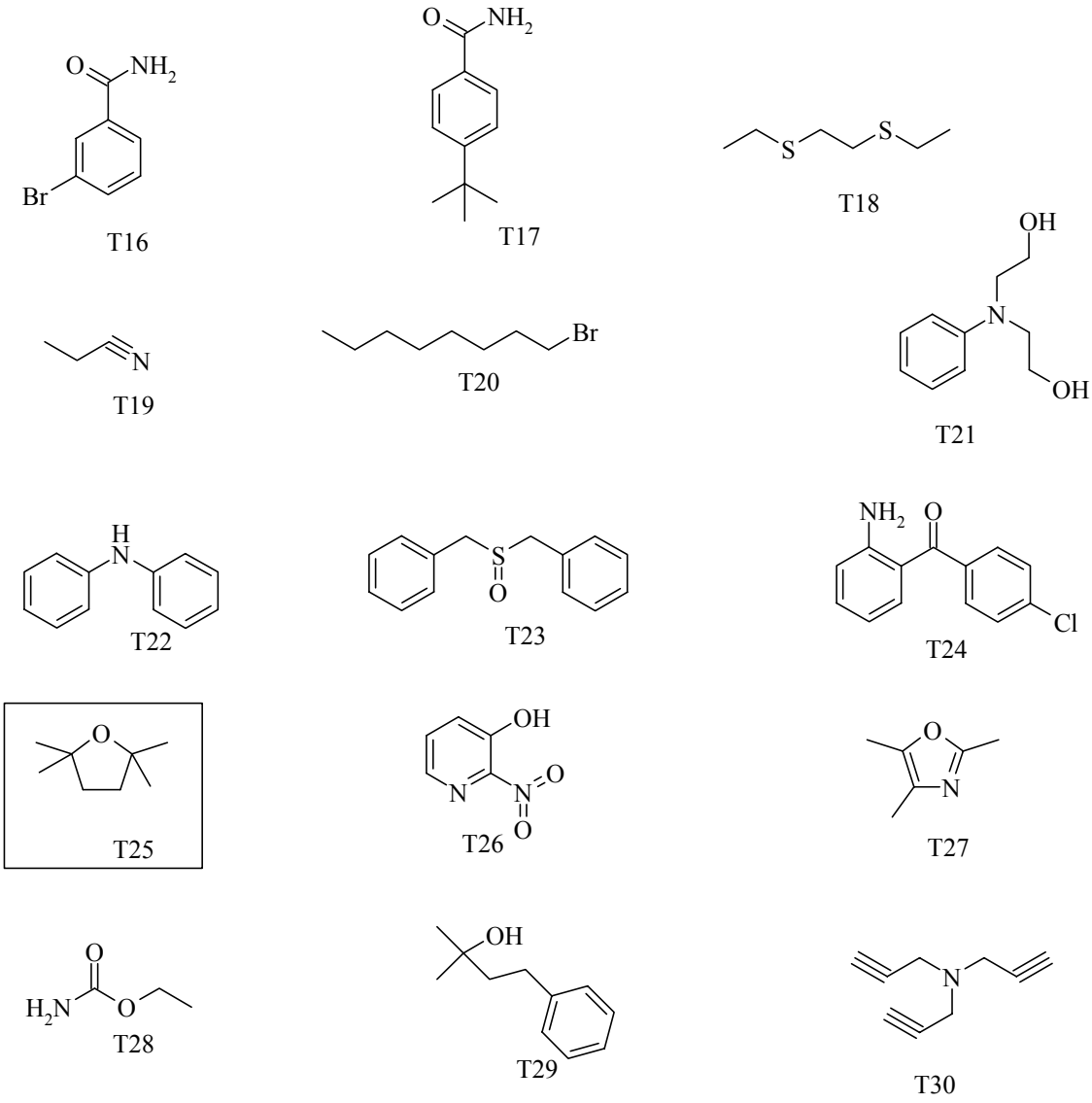

Figure 2. Compounds in the test set

For all the 220 molecules we have calculated the atomic descriptors and selected those useful for the model. 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 hydrogen 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 observing the fish behavior 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 syndromes 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 syndrome allocation. We generally accept the Ivanciuc 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% 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 11 or 8 misclassifications (T8-T10, T12, T15, T19, T22, T25-T27, T30), or (T8-T10, T15, T22, T25-T26, T30). It is remarkable that many misclassifications present in HQ^+ are also present in Helc.

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 (T8-T10, T12, T15, T19, T22, T25-T27; T8-T10, T14-T15, T22, T25-T26, 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 1. Classical-NN representative results

		training	misclassified	validation	misclassified
Helc+Elcdif	class 1	101	1	12	0
	class 2	64	4	8	0
test	class 1	18	7		
	class 2	1	4		
rand	class 1	0	84	0	10
	class 2	86	0	10	0
HQ+Elcdif	class 1	96	1	17	0
	class 2	69	4	3	0
test	class 1	21	4		
	class 2	1	4		
rand	class 1	87	0	7	0
	class 2	0	83	0	13

Table 2. Tree results

		training	misclassified	nodes	levels
Helc+Elcdif	class 1	110	4	4	4
	class 2	75	1		
test	class 1	19	6	3	
	class 2	1	4		
rand	class 1	69	45	10	11
	class 2	51	25		
HQ+Elcdif	class 1	110	4	3	3
	class 2	74	2		
test	class 1	21	4	3	
	class 2	0	5		
rand	class 1	70	44	10	8
	class 2	46	30		

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:

Using Helc

1. 1-60; *54* misclassified
2. 61-114
3. 115-130; 156-157
4. 131-190; *164* and *181-183* misclassified

Using HQ^+

1. 1-32, 56, 60
2. 33-114; *54* and *75* misclassified
3. 115-190; *164* and *181-183* misclassified

If we translate the subsets into chemical terms, we have:

Using Helc

1. alcohols, ketones, esters, ethers; *diphenyl ether* misclassified
2. halides, hydrocarbons
3. nitro compounds, pyridine and quinoline
4. phenols, anilines; *N,N-dimethyl aniline* and *fluoro anilines* misclassified

Using HQ^+

1. alcohols, furan, 2-hydroxy-4-methoxy acetophenone
2. ketones, esters, ethers, halides, hydrocarbons; *diphenyl ether* and *trichloroethene* misclassified
3. nitro compounds; pyridine and quinoline, phenols, anilines; *N,N-dimethyl aniline* and *fluoro anilines* misclassified

For what the test set is concerned we have the following classification:

Using Helc

1. 1-7, 11, 13, 16-17, 21, 23-24, 28-29
2. 14, 18, 20, 30; *8-10, 12, 15, 19, 22, 25-27*, misclassified

Using HQ^+

1. 1-7, 12-13, 18-20, 23, 27
2. 11, 16-17, 21, 24, 28-29; *8-10, 14-15, 22, 25-26, 30*, misclassified

If we translate the misclassified compounds in chemical terms, we have:

Using Helc

diphenyl sulfoxide, alkyl nitriles, primary amine, nitro pyridinol, oxazole, triphenyl phosphine oxide, tributoxy-ethoxy phosphate, 4-trifluoromethyl-3-nitro phenol, 2,2,5,5-tetramethyl tetrahydrofurane

Using HQ^+

diphenyl sulfoxide, pyridinyl nitrile, primary amine, nitro pyridinol, tripropargyl amine, triphenyl phosphine oxide, tributoxy-ethoxy phosphate, 4-trifluoromethyl-3-nitro phenol, 2,2,5,5-tetramethyl tetrahydrofurane

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).

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 behaviour. 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. Let's do some examples.

1,1,1-trichloroethane causes a class 2 syndrome;

N,N-dimethylaniline causes a class 1 syndrome;

4-ethylaniline causes a class 1 syndrome;

4-chlorophenol, 4-methoxyphenol, and pyridine, cause a class 3 syndrome.

It is evident that the fish reaction to chemicals is more complex than that predicted by models. The test set is even more complicated. We find ~10/30 misclassifications, but we must take into consideration that here the class assignment is done following Russom et al. [2] We have difficulties in estimating the relative value of experimental fish behaviour and of calculated prediction. The experimental results have different levels of confidence, as reported by Russom et al., [2] but they are real effects. On the other hand, calculated predictions 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 toxicity, but we must be very careful throughout our analysis.

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

The descriptor values and class assignment are reported in the following Tables.

Table 3. Structure of the chemical compounds in training set, theoretical descriptors (Elcdfi, Helc, HQ^+) and mechanism of toxic action (nonpolar, +1; polar, –1; experimental, Exp; prediction, Pre)

No	Compound	Elcdif	Helc	HQ^+		C-NN/ Helc	C-NN/ HQ	Tree/ Helc	Tree/ HQ	SVM
					Exp	Pre	Pre	Pre	Pre	Pre
1	methanol	2.56	1	242	+1	+1	+1	+1	+1	+1
2	ethanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
3	1–propanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
4	2–propanol	2.56	0.96	243	+1	+1	+1	+1	+1	+1
5	1–butanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
6	2–butanol	2.57	0.96	243	+1	+1	+1	+1	+1	+1
7	isobutanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
8	*tert*–butyl alcohol	2.56	0.94	243	+1	+1	+1	+1	+1	+1
9	1–pentanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
10	3–pentanol	2.56	0.96	243	+1	+1	+1	+1	+1	+1
11	1–hexanol	2.56	1.01	242	+1	+1	+1	+1	+1	+1
12	1–heptanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
13	1–octanol	2.56	0.98	244	+1	+1	+1	+1	+1	+1
14	1–nonanol	2.56	0.98	244	+1	+1	+1	+1	+1	+1
15	1–decanol	2.56	0.98	244	+1	+1	+1	+1	+1	+1
16	1–undecanol	2.56	0.98	244	+1	+1	+1	+1	+1	+1
17	1–dodecanol	2.56	0.98	244	+1	+1	+1	+1	+1	+1
18	1,2–ethanediol	2.56	0.98	243	+1	+1	+1	+1	+1	+1
19	1,3–propenediol	2.56	0.98	243	+1	+1	+1	+1	+1	+1
20	2–methyl–2,4–pentanediol	2.56	0.96	243	+1	+1	+1	+1	+1	+1
21	3–furanmethanol	2.56	1.02	243	+1	+1	+1	+1	+1	**–1**
22	cyclohexanol	2.56	0.96	243	+1	+1	+1	+1	+1	+1
23	2,2,2–trichloroethanol	2.57	0.98	242	+1	+1	+1	+1	+1	**–1**
24	butyldigol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
25	diethyleneglycol	2.56	0.98	243	+1	+1	+1	+1	+1	+1
26	triethyleneglycol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
27	2–methoxyethanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
28	2–ethoxyethanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
29	2–isopropoxyethanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
30	2–butoxyethanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
31	2–(2–ethoxyethoxy) ethanol	2.56	0.98	242	+1	+1	+1	+1	+1	+1
32	2–phenoxyethanol	2.56	0.98	242	+1	+1	+1	+1	+1	**–1**
33	acetone	2.18	1.21	19	+1	+1	+1	+1	+1	+1
34	2–propanone	2.18	1.21	20	+1	+1	+1	+1	+1	+1
35	2–butanone	2.18	1.21	20	+1	+1	+1	+1	+1	+1
36	3–pentanone	2.18	1.21	20	+1	+1	+1	+1	+1	+1
37	2–octanone	2.18	1.21	20	+1	+1	+1	+1	+1	+1
38	5–nonanone	2.18	1.16	20	+1	+1	+1	+1	+1	+1
39	2–decanone	2.18	1.21	20	+1	+1	+1	+1	+1	+1
40	3–methyl–2–butanone	2.18	1.21	21	+1	+1	+1	+1	+1	+1
41	6–methyl–5–hepten–2–one	2.18	1.21	37	+1	+1	+1	+1	+1	+1
42	2,3,4–trimethoxy acetophenone	2.19	1.14	36	+1	+1	+1	+1	+1	+1
43	acetophenone	2.18	1.12	37	+1	+1	+1	+1	+1	+1
44	3,3–dimethyl–2–butanone	2.18	1.21	19	+1	+1	+1	+1	+1	+1
45	4–methyl–2–pentanone	2.18	1.21	20	+1	+1	+1	+1	+1	+1
46	benzophenone	2.14	1.03	37	+1	+1	+1	+1	+1	+1
47	2,4–dichloro acetophenone	2.18	1.12	36	+1	+1	+1	+1	+1	**–1**

Table 3. (Continued)

No	Compound	Elcdif	Helc	HQ+		C-NN/ Helc	C-NN/ HQ	Tree/ Helc	Tree/ HQ	SVM
					Exp	Pre	Pre	Pre	Pre	Pre
48	cyclohexanone	2.18	1.21	20	+1	+1	+1	+1	+1	+1
49	ethyl acetate	2.18	1.11	28	+1	+1	+1	+1	+1	+1
50	diethyl ether	2.18	0.95	27	+1	+1	+1	+1	+1	+1
51	diiso–propyl ether	2.18	0.91	29	+1	+1	+1	+1	+1	+1
52	dibutyl ether	2.19	0.94	27	+1	+1	+1	+1	+1	+1
53	dipentyl ether	2.19	0.94	27	+1	+1	+1	+1	+1	+1
54	diphenyl ether	2.22	0.74	40	+1	**–1**	**–1**	**–1**	**–1**	+1
55	*tert*–butylmethyl ether	2.19	0.95	26	+1	+1	+1	+1	+1	+1
56	furan	2.15	1.03	62	+1	+1	+1	+1	+1	+1
57	tetrahydrofuran	2.19	0.94	27	+1	+1	+1	+1	+1	+1
58	2,6–dimethoxytoluene	2.19	0.96	36	+1	+1	+1	+1	+1	+1
59	1,4–dimethoxybenzene	2.19	0.96	36	+1	+1	+1	+1	+1	+1
60	2–hydroxy–4–methoxy acetophenone	2.57	0.96	243	+1	+1	+1	+1	+1	**–1**
61	dichloromethane	2.19	0.47	25	+1	+1	+1	+1	+1	+1
62	chloroform	2.19	0.46	29	+1	+1	+1	+1	+1	–1
63	tetrachloromethane	0	0.47	0	+1	+1	+1	+1	+1	+1
64	1,1–dichloroethane	2.19	0.44	26	+1	+1	+1	+1	+1	+1
65	1,2–dichloroethane	2.18	0.46	22	+1	+1	+1	+1	+1	+1
66	1,1,1–trichloroethane	2.18	0.42	18	+1	+1	+1	+1	+1	+1
67	1,1,2–trichloroethane	2.18	0.46	26	+1	+1	+1	+1	+1	+1
68	1,1,2,2–tetrachloroethane	2.19	0.44	26	+1	+1	+1	+1	+1	**–1**
69	pentachloroethane	2.18	0.44	26	+1	+1	+1	+1	+1	**–1**
70	hexachloroethane	0	0.42	0	+1	+1	+1	+1	+1	+1
71	1,2–dichloropropane	2.18	0.46	23	+1	+1	+1	+1	+1	+1
72	1,3–dichloropropane	2.18	0.46	22	+1	+1	+1	+1	+1	+1
73	1,2,3–trichloropropane	2.18	0.46	23	+1	+1	+1	+1	+1	+1
74	1–chlorobutane	2.18	0.46	22	+1	+1	+1	+1	+1	+1
75	trichloroethene	2.13	0.35	40	+1	+1	+1	+1	**–1**	+1
76	tetrachloroethene	0	0.28	0	+1	+1	+1	+1	+1	+1
77	hexachlorobutadiene	0	0.3	0	+1	+1	+1	+1	+1	+1
78	lindane	2.18	0.36	23	+1	+1	+1	+1	+1	+1
79	chlorobenzene	2.14	0.3	34	+1	+1	+1	+1	+1	+1
80	1,2–dichlorobenzene	2.14	0.3	34	+1	+1	+1	+1	+1	+1
81	1,3–dichlorobenzene	2.14	0.3	35	+1	+1	+1	+1	+1	+1
82	1,4–dichlorobenzene	2.14	0.3	34	+1	+1	+1	+1	+1	+1
83	1,2,3–trichlorobenzene	2.14	0.3	34	+1	+1	+1	+1	+1	+1
84	1,2,4–trichlorobenzene	2.14	0.3	35	+1	+1	+1	+1	+1	+1
85	1,3,5–trichlorobenzene	2.14	0.3	35	+1	+1	+1	+1	+1	+1
86	1,2,3,4–tetrachloro benzene	2.14	0.29	34	+1	+1	+1	+1	+1	+1
87	1,2,3,5–tetrachloro benzene	2.14	0.29	35	+1	+1	+1	+1	+1	+1
88	1,2,4,5–tetrachloro benzene	2.14	0.29	35	+1	+1	+1	+1	+1	+1
89	3–chlorotoluene	2.18	0.26	36	+1	+1	+1	+1	+1	+1
90	4–chlorotoluene	2.18	0.3	34	+1	+1	+1	+1	+1	+1
91	2,4–dichlorotoluene	2.18	0.3	35	+1	+1	+1	+1	+1	+1
92	2,4,5–trichlorotoluene	2.18	0.3	35	+1	+1	+1	+1	+1	+1
93	3,4–dichlorotoluene	2.18	0.3	34	+1	+1	+1	+1	+1	+1
94	pentachlorobenzene	2.14	0.3	35	+1	+1	+1	+1	+1	+1
95	2–chloronaphthalene	2.14	0.3	35	+1	+1	+1	+1	+1	+1
96	hexane	2.18	0.02	19	+1	+1	+1	+1	+1	+1
97	octane	2.18	0.02	19	+1	+1	+1	+1	+1	+1

Table 3. (Continued)

No	Compound	Elcdif	Helc	HQ+		C-NN/ Helc	C-NN/ HQ	Tree/ Helc	Tree/ HQ	SVM
					Exp	Pre	Pre	Pre	Pre	Pre
98	decane	2.18	0.02	19	+1	+1	+1	+1	+1	+1
99	benzene	2.14	0	34	+1	+1	+1	+1	+1	+1
100	toluene	2.18	0.24	34	+1	+1	+1	+1	+1	+1
101	*o*–xylene	2.18	0.22	34	+1	+1	+1	+1	+1	+1
102	*m*–xylene	2.18	0.23	34	+1	+1	+1	+1	+1	+1
103	*p*–xylene	2.18	0.23	34	+1	+1	+1	+1	+1	+1
104	1,2,4–trimethylbenzene	2.18	0.22	34	+1	+1	+1	+1	+1	+1
105	1,3,5–trimethylbenzene	2.18	0.23	34	+1	+1	+1	+1	+1	+1
106	1,2,4,5–tetramethyl benzene	2.18	0.23	34	+1	+1	+1	+1	+1	+1
107	ethylbenzene	2.18	0.22	34	+1	+1	+1	+1	+1	+1
108	cumene	2.18	0.19	34	+1	+1	+1	+1	+1	+1
109	1–methylnaphthalene	2.18	0.23	34	+1	+1	+1	+1	+1	+1
110	2–methylnaphthalene	2.18	0.23	34	+1	+1	+1	+1	+1	+1
111	biphenyl	2.14	0.07	34	+1	+1	+1	+1	+1	+1
112	cyclopentane	2.18	0	19	+1	+1	+1	+1	+1	+1
113	cyclohexane	2.18	0	19	+1	+1	+1	+1	+1	+1
114	methylcyclohexane	2.18	0.03	20	+1	+1	+1	+1	+1	+1
115	nitrobenzene	2.14	0.82	39	–1	–1	–1	–1	–1	–1
116	2–nitrotoluene	2.18	0.83	39	–1	–1	–1	–1	–1	–1
117	3–nitrotoluene	2.18	0.82	39	–1	–1	–1	–1	–1	–1
118	4–nitrotoluene	2.18	0.82	39	–1	–1	–1	–1	–1	–1
119	2,3–dimethylnitrobenzene	2.19	0.83	39	–1	–1	–1	–1	–1	–1
120	3,4–dimethylnitrobenzene	2.18	0.82	39	–1	–1	–1	–1	–1	–1
121	2–chloronitrobenzene	2.14	0.82	39	–1	–1	–1	–1	–1	–1
122	3–chloronitrobenzene	2.15	0.82	39	–1	–1	–1	–1	–1	–1
123	4–chloronitrobenzene	2.15	0.82	39	–1	–1	–1	–1	–1	–1
124	2,3–dichloronitrobenzene	2.15	0.82	39	–1	–1	–1	–1	–1	–1
125	2,4–dichloronitrobenzene	2.15	0.82	39	–1	–1	–1	–1	–1	–1
126	2,5–dichloronitrobenzene	2.15	0.82	39	–1	–1	–1	–1	–1	–1
127	3,5–dichloronitrobenzene	2.15	0.82	39	–1	–1	–1	–1	–1	–1
128	2–chloro–6–nitrotoluene	2.19	0.83	39	–1	–1	–1	–1	–1	–1
129	4–chloro–2–nitrotoluene	2.19	0.83	39	–1	–1	–1	–1	–1	–1
130	4–chloro–3–nitrotoluene	2.18	0.82	39	–1	–1	–1	–1	–1	–1
131	phenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
132	2–methylphenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
133	3–methylphenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
134	4–methylphenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
135	2,4–dimethylphenol	2.57	0.86	242	–1	–1	–1	–1	–1	–1
136	2,6–dimethylphenol	2.57	0.87	243	–1	–1	–1	–1	–1	–1
137	3,4–dimethylphenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
138	2,3,6–trimethylphenol	2.57	0.87	243	–1	–1	–1	–1	–1	–1
139	2,4,6–trimethylphenol	2.57	0.87	243	–1	–1	–1	–1	–1	–1
140	4–ethylphenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
141	4–propylphenol	2.56	0.86	244	–1	–1	–1	–1	–1	–1
142	4–*n*–butylphenol	2.56	0.86	244	–1	–1	–1	–1	–1	–1
143	4–*tert*–butylphenol	2.56	0.86	244	–1	–1	–1	–1	–1	–1
144	2–*tert*–butyl–4–methylphenol	2.57	0.87	244	–1	–1	–1	–1	–1	–1
145	4–*n*–pentylphenol	2.56	0.86	244	–1	–1	–1	–1	–1	–1
146	4–*tert*–pentylphenol	2.56	0.86	244	–1	–1	–1	–1	–1	–1
147	2–allylphenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1

Table 3. (Continued)

No	Compound	Elcdif	Helc	HQ+		C-NN/ Helc	C-NN/ HQ	Tree/ Helc	Tree/ HQ	SVM
					Exp	Pre	Pre	Pre	Pre	Pre
148	2–phenylphenol	2.56	0.87	241	–1	–1	–1	–1	–1	–1
149	1–naphthol	2.57	0.86	242	–1	–1	–1	–1	–1	–1
150	4–chlorophenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
151	4–chloro–3–methylphenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
152	4–chloro–3,5–dimethyl phenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
153	3–methoxyphenol	2.56	0.86	242	–1	–1	–1	–1	–1	–1
154	4–methoxyphenol	2.56	0.76	242	–1	–1	–1	–1	–1	–1
155	4–phenoxyphenol	2.56	0.76	242	–1	–1	–1	–1	–1	–1
156	pyridine	2.15	0.68	40	–1	–1	–1	–1	–1	**+1**
157	quinoline	2.15	0.67	41	–1	–1	–1	–1	–1	**+1**
158	aniline	2.35	0.27	100	–1	–1	–1	–1	–1	–1
159	2–methylaniline	2.36	0.27	100	–1	–1	–1	–1	–1	–1
160	3–methylaniline	2.36	0.26	100	–1	–1	–1	–1	–1	–1
161	4–methylaniline	2.35	0.27	100	–1	–1	–1	–1	–1	–1
162	2,3–dimethylaniline	2.35	0.27	100	–1	–1	–1	–1	–1	–1
163	3,4–dimethylaniline	2.35	0.27	100	–1	–1	–1	–1	–1	–1
164	*N*,*N*–dimethylaniline	2.18	0.3	34	–1	**+1**	**+1**	**+1**	**+1**	**+1**
165	2–ethylaniline	2.35	0.27	100	–1	–1	–1	–1	–1	–1
166	3–ethylaniline	2.36	0.27	100	–1	–1	–1	–1	–1	–1
167	4–ethylaniline	2.36	0.26	100	–1	–1	–1	–1	–1	–1
168	4–butylaniline	2.35	0.26	100	–1	–1	–1	–1	–1	–1
169	2,6–diisopropylaniline	2.36	0.28	100	–1	–1	–1	–1	–1	–1
170	2–chloroaniline	2.36	0.27	100	–1	–1	–1	–1	–1	–1
171	3–chloroaniline	2.35	0.27	100	–1	–1	–1	–1	–1	–1
172	4–chloroaniline	2.36	0.27	100	–1	–1	–1	–1	–1	–1
173	2,4–dichloroaniline	2.36	0.27	100	–1	–1	–1	–1	–1	–1
174	2,5–dichloroaniline	2.36	0.27	100	–1	–1	–1	–1	–1	–1
175	3,4–dichloroaniline	2.35	0.26	100	–1	–1	–1	–1	–1	–1
176	3,5–dichloroaniline	2.35	0.27	100	–1	–1	–1	–1	–1	–1
177	2,3,4–trichloroaniline	2.35	0.27	100	–1	–1	–1	–1	–1	–1
178	2,3,6–trichloroaniline	2.36	0.26	100	–1	–1	–1	–1	–1	–1
179	2,4,5–trichloroaniline	2.35	0.26	100	–1	–1	–1	–1	–1	–1
180	4–bromoaniline	2.35	0.26	99	–1	–1	–1	–1	–1	–1
181	α,α,α,4–tetrafluoro–3–methylaniline	2.35	1.18	101	–1	**+1**	**+1**	**+1**	**+1**	–1
182	α,α,α,4–tetrafluoro–2–methylaniline	2.36	1.18	101	–1	**+1**	**+1**	**+1**	**+1**	–1
183	pentafluoroaniline	2.36	1.17	104	–1	**+1**	**+1**	**+1**	**+1**	–1
184	3–benzyloxyaniline	2.35	0.72	101	–1	–1	–1	–1	–1	–1
185	4–hexyloxyaniline	2.35	0.72	100	–1	–1	–1	–1	–1	–1
186	2–nitroaniline	2.36	0.82	103	–1	–1	–1	–1	–1	–1
187	3–nitroaniline	2.35	0.82	100	–1	–1	–1	–1	–1	–1
188	4–nitroaniline	2.35	0.82	101	–1	–1	–1	–1	–1	**+1**
189	2–chloro–4–nitroaniline	2.35	0.82	101	–1	–1	–1	–1	–1	–1
190	4–ethoxy–2–nitroaniline	2.35	0.82	103	–1	–1	–1	–1	–1	–1

Table 4. Structure of the chemical compounds in test set, theoretical descriptors (Elcdfi, Helc, HQ+) and mechanism of toxic action (nonpolar, +1; polar, –1; experimental, Exp; prediction, Pre)

No	Compound	Elcdif	Helc	HQ+		C-NN/ Helc	C-NN/ HQ	Tree/ Helc	Tree/ HQ
					Exp	Pre	Pre	Pre	Pre
T1	diethyl benzyl phosphonate	1.04	2.22	38	+1	+1	+1	+1	+1
T2	azinphos-methyl	1.03	2.22	38	ACI	+1	+1	+1	+1
T3	malathion	1.1	2.22	29	ACI	+1	+1	+1	+1
T4	disulfoton	0.97	2.22	28	ACI	+1	+1	+1	+1
T5	terbufos	0.97	2.22	28	ACI	+1	+1	+1	+1
T6	carbophenothion	0.97	2.22	36	ACI	+1	+1	+1	+1
T7	tributyl phosphate	1	2.22	29	+1	+1	+1	+1	+1
T8	tris(2-butoxyethyl) phosphate	1	2.22	29	–1	**+1**	**+1**	**+1**	**+1**
T9	triphenyl phosphine oxide	1.1	2.22	39	–1	**+1**	**+1**	**+1**	**+1**
T10	diphenyl sulfoxide	0.85	2.22	40	+1	**–1**	**–1**	**–1**	**–1**
T11	phenobarbital	1.14	2.36	118	+1	+1	+1	+1	+1
T12	octyl cyanide	0.7	2.18	22	+1	**–1**	+1	**–1**	+1
T13	2,6-dimethoxy toluene	0.66	2.19	36	+1	+1	+1	+1	+1
T14	2-cyano pyridine	0.65	2.14	40	–1	–1	–1	–1	**+1**
T15	3-trifluoromethyl-4-nitro phenol	1.1	2.5	233	–1	**+1**	**+1**	**+1**	**+1**
T16	m-bromo benzamide	1.04	2.35	105	+1	+1	+1	+1	+1
T17	p-tbutyl benzamide	1.04	2.35	105	+1	+1	+1	+1	+1
T18	3,6-dithia octane	0.11	2.18	19	+1	+1	+1	+1	+1
T19	proprionitrile	0.7	2.18	21	+1	**–1**	+1	**–1**	+1
T20	1-bromo octane	0.24	2.18	20	+1	+1	+1	+1	+1
T21	N-phenyl diethanol amine	0.97	2.56	242	+1	+1	+1	+1	+1
T22	diphenyl amine	0.37	2.39	111	+1	**–1**	**–1**	**–1**	**–1**
T23	dibenzyl sulfoxide	1.14	2.14	34	+1	+1	+1	+1	+1
T24	2-amino-4'-chloro benzophenone	1.14	2.39	105	+1	+1	+1	+1	+1
T25	2,2,5,5-tetramethyl tetrahydrofuran	0.9	2.18	18	–1	**+1**	**+1**	**+1**	**+1**
T26	3-hydroxy-2-nitro pyridine	0.76	2.51	236	+1	**–1**	**–1**	**–1**	**–1**
T27	2,4,5-trimethyl oxazole	0.78	2.18	19	+1	**–1**	+1	**–1**	+1
T28	urethane	1.04	2.35	102	+1	+1	+1	+1	+1
T29	benzyl tert buthanol	0.94	2.56	245	+1	+1	+1	+1	+1
T30	tripropargyl amine	0.52	2.18	70	+1	**–1**	**–1**	+1	**–1**

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.
[6] A. Ben–Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, Support Vector Clustering, *J. Machine Learning Res.* **2001**, *2*, 125–137.
[7] 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**.