

## QSAR Model for Predicting Pesticide Aquatic Toxicity

Paolo Mazzatorta, Martin Smiesko, Elena Lo Piparo,\* and Emilio Benfenati

Istituto di Ricerche Farmacologiche “Mario Negri” Milano, Via Eritrea, 62, 20157 Milano, Italy

Received June 17, 2005

A hierarchical QSAR approach was applied for the prediction of acute aquatic toxicity. Chemical structures were encoded into molecular descriptors by an automated, seamless procedure available within the OpenMolGRID system. Finally, various linear and nonlinear regression techniques were used to obtain stable and thoroughly validated QSARs. The final model was developed by a counterpropagation neural network coupled with genetic algorithms for variable selection. The proposed QSAR is consistent with McFarland's principle for biological activity and makes use of seven molecular descriptors, namely HACA-2, HOMO–LUMO energy gap, Kier and Hall index, HA dependent HDSA-1, BETA polarizability, FHBCA fractional HBSA, and LogP. The model was extensively tested by the test set ( $R^2 = 0.79$ ), the y-scrambling test, and sensitivity/stability tests.

### 1. INTRODUCTION

Pesticides occupy a unique position among the numerous chemicals that man encounters daily, in that they are deliberately added to the environment to kill some form of life. Ideally, their injurious action would be highly specific for undesirable target organisms and nontoxic to desirable, nontarget organisms. The goodness of a pesticide should be based on its specificity but often relies on its rational usage and dosage. In fact, most of the chemicals that are used as pesticides are not highly selective but are toxic to many nontarget species, including man, and other desirable forms of life that cohabit the environment. This characteristic, together with their massive employment (every year an estimated 2.5 million tons of pesticides are applied to agricultural crops worldwide), makes pesticides among the most dangerous chemicals in the world.<sup>1,2</sup>

An increasing number of environmental effects of pesticide applications are now being taken into account by regulatory bodies, leading to increased restrictions on their use or even bans. In particular, the acute toxicity test with fish is a key item in current European and American guidelines for testing chemicals.<sup>3–5</sup> Its aim is to assess the risk to similar species in natural environments, as an aid in the determination of water quality criteria for regulatory purposes. Data on one cold and one warm freshwater species are generally required. The rainbow trout, *Oncorhynchus mykiss*, and bluegill sunfish, *Lepomis macrochirus*, are the preferred species to meet this requirement since they are sensitive, and there is a large database characterizing the response to environmental contaminants.

On the other hand, the cost of in vivo testing, i.e., in living animals, is prohibitive and weighs heavily on the final price of chemicals. In a recent white paper,<sup>6</sup> the European Commission estimated that “the testing of the approximately 30,000 existing substances would result in total costs of about 2.1 billion (euro), over the next 11 years until 2012”. Besides

economic constraints, ethical considerations and public pressure work to reduce tests on animals.<sup>7</sup> Thus, more convenient and efficient methods to predict biological activity from structural information are needed. Moreover, the use of validated methods is encouraged by the EU and U.S.A.<sup>8</sup>

In this framework, QSAR (quantitative structure–activity relationship) approaches are challenging methods to cover many knowledge gaps. The basic assumption of QSAR is that a quantitative relationship exists between the molecular structure of compounds and their biological, chemical, and physical properties, and therefore knowing such a relationship it is possible to predict the property of a given chemical entity from its structure, without any direct experimental measurement. A comprehensive description of the QSAR approach in toxicology was published by Schultz and Cronin.<sup>9</sup>

### 2. MATERIAL AND METHOD

**Toxicity Data.** The data set was collected within the EU funded project DEMETRA. Toxicological data refer to the acute toxicity LC<sub>50</sub> 96h [ppm] exposure for the rainbow trout and were extracted from the U.S. EPA-OPP (EPA-Office of Pesticides Programs), SEEM (produced by the International Center for Pesticides and Health Risk Prevention), and BBA (Federal Biological Research Center for Agriculture and Forestry) ecotoxicological database for aquatic data. To ensure high quality for QSAR development, criteria for consistency and reliability of the data were applied, as described in ref 10. Briefly see the following:

- data obtained according to a standardized procedure such as OECD<sup>3</sup> or EPA guidelines;<sup>5</sup>
- data obtained according to Good Laboratory Practice (GLP);
- availability of ancillary data such as purity, year of the study, uncertainty of the experimental result, and other statistical parameters.

At last 282 chemicals, spanning a wide range of chemical classes, were associated with reliable experimental values. Data are confidential since related to the process of approval of pesticides. However, data of pesticides on the EU market

\*Corresponding author phone: +39-02-39014420; fax: +39-02-39001916; e-mail: lopiparo@marionegri.it.

are available, and data from the SEEM project have been incorporated in the database which will be published by the EC within the project RED. As a common good practice in ecotoxicological QSARs, data were then transformed and modeled as  $\text{Log}_{10}(1/\text{LC}_{50})$  [50% lethal concentration expressed in mmol/L].

**Chemical Descriptors.** Chemical descriptors were obtained using the OpenMolGRID system. The OpenMolGRID project is developing tools for secure and seamless access to distributed information and computational methods relevant for molecular engineering. A full description of the system can be found in refs 11 and 12. We used the following workflow:

- 2D structure sketches (connectivity formulas) are input into the OpenMolGRID system;
- structures are converted to three dimensions using the computer program MOLGEO 1.0 [algorithm settings: distance geometry; tolerance: 3; time limit: 10; add hydrogens: ON];
- 3D structures are optimized by semiempirical methods implemented in MOPAC 7.05 using the gradient criterion [keywords: AM1 T=3600 NOINTER MMOK GNORM=0.1 EF];
- single point calculations of thermodynamic properties at optimized geometries are done using MOPAC 7.05 [keywords: AM1 VECTORS BONDS PI POLAR PRECISE ENPART MMOK 1SCF];
- optimized 3D structures of the compounds with additional thermodynamic output files for each structure are used as input for CODESSA PRO software, which calculates a large pool of constitutional, geometric, topological, electrostatic, surface area, quantum-chemical, molecular orbital, and thermodynamic descriptors.

The logarithms of octanol/water partition coefficient  $\text{LogP}$  were calculated by KowWin1 and added to the data set because of its relevance in predicting aquatic acute toxicity. In total 1048 descriptors were calculated.

The whole trout data set consisted of 282 compounds. Initially, eight compounds were excluded from the data set, because their particular conformations did not allow automatic modeling. These compounds were later manually inspected and remodeled.

**Statistical Techniques.** Various statistical techniques were used for extracting information and deriving predictive models:

- Multilinear regression (MLR): The purpose of MLR is to fit a linear model that minimize the difference, or error, between predicted and actual values.
- Partial least-squares (PLS):<sup>13</sup> This is based on a linear transformation of the descriptor space, producing a new variable space based on a small number of orthogonal factors (latent variables), so there is no correlation. The NIPALS algorithm was used.
- Back-propagation neural networks (BPNN): These are among the most popular neural network architectures used nowadays in chemometrics. BPNNs are made up of neurons organized in layers and connected through weights. During the learning phase the weights of the BPNN are modified so that the response to a given input (descriptor) is similar to the target (activity). A BPNN with three layers and an appropriate number of hidden neurons (in the middle layer) can fit any function to a given accuracy.<sup>14</sup> For this study we

used a three-layer network {tansig; tansig; pureline} with 10-15-1 neurons, 100 training epochs, *traingdx* learning function, *mse* performance function, 0.01 learning rate, and momentum constant 0.95. For details about parameters refer to the Neural Network Toolbox for Matlab.<sup>15</sup>

• A combination of genetic algorithm and counterpropagation neural network (GA/CPANN): GA is a stochastic global search method that mimics the natural biological evolution.<sup>16</sup> The architecture of a CPANN consists of two layers of neurons, the input or Kohonen layer and the output layer.<sup>17,18</sup> The input layer is unsupervised, and the learning procedure is the same as in the Kohonen networks; the output layer is supervised and reflects the topological position that inputs find in the Kohonen layer. GA explores the descriptor hyperspace for selection of variables, and CPANN derives the fitness score. Details of the procedure are given in ref 19. For this study, three parallel populations of eight individuals evolved through 200 generations, to optimize a 12-by-12 network, with 100 training epochs.

All the calculations were performed using Matlab 7 (R14) (The MathWorks, Natick, MA).

To measure performances we used determination coefficient ( $R^2$ ), and root-mean-squared error (rmse), calculated as follows:

$$r = y - \hat{y}$$

$$\text{rmse} = \sqrt{\left(\frac{\sum r^2}{n}\right)}$$

$$\text{sse} = \sum r^2$$

$$\text{ssr} = \sum (y - \bar{y})^2$$

$$R^2 = 1 - \frac{\text{sse}}{\text{ssr}}$$

where  $y$  is the experimental value,  $\hat{y}$  is the predicted value,  $n$  is the number of data points, and  $\bar{y}$  is the mean of the experimental values.

**Preprocessing.** To reduce the risk of chance correlation<sup>20,21</sup> and overfitting of data,<sup>22</sup> the data set is analyzed using filters to remove descriptors with either small variance or no unique information. The data set is preprocessed in columns and rows in order to eliminate constant variables, empty values, and intercorrelated descriptors. Thus

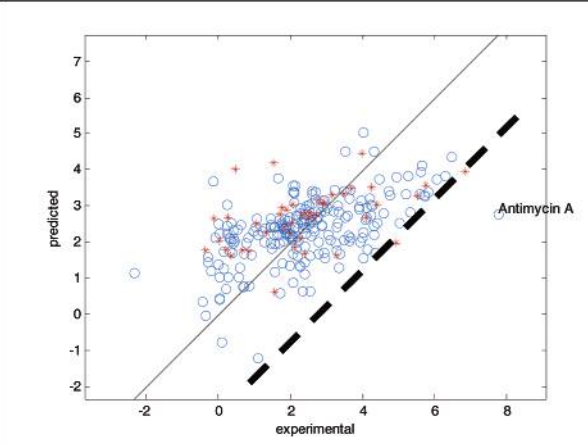
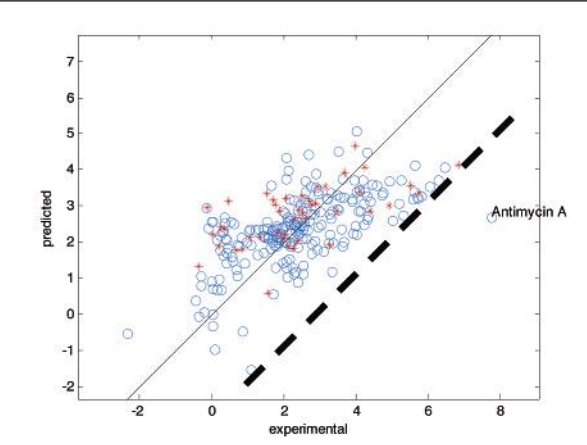
- 14 variables with standard deviation equal to zero (constant variables) were eliminated;
- 8 chemicals did not have enough information (more than 80% missing values)—these compounds contain metals such as arsenic or tin, for which semiempirical calculation cannot be done;
- 669 variables still had missing values;
- 46 variables had an intercorrelation equal to one.

A total of 729 variables and 8 chemicals were discarded from the study leaving a data set with 274 chemicals and 319 descriptors.

The importance of preprocess data is well-known in QSAR analysis;<sup>23</sup> therefore, the data set was scaled so to have zero mean and unity standard deviation.

The models generated were thoroughly validated by appropriate statistical procedures, including the prediction

Table 1. Expert Modeling Results<sup>a</sup>

Tox = f(HOMO,logP)			Tox = f(LUMO,logP)		
					
$-\log(LC_{50}) = 0.0183HOMO + 0.4045 \log P + 1.2419$			$-\log(LC_{50}) = -0.446LUMO + 0.3683 \log P + 0.9542$		
	$R^2$	rmse		$R^2$	rmse
Training	0.32	1.282	Training	0.397	1.204
Test	0.19	1.447	Test	0.311	1.331

<sup>a</sup> Blue circles (○) are chemicals in the training set, red asterisks (\*) are chemicals in the test set, and dashed lines depict baseline toxicity.

of an external test set. The data set was split into two separate sets: (1) training set [222 data points], used for training the models, and (2) test set [52 data points], used for testing the real predictive ability of the models.

Ideally, the data set should be split in order to be representative of a well-defined and significant chemical space.<sup>24</sup> But this procedure would imply the a priori knowledge and/or the selection of variables relevant to the problem. Moreover, the chemical space describing the data set is not unique, but the use of different software, e.g. Dragon, Sybyl, etc., leads to a different chemical space. For these reasons the data set was split into the previously defined sets, randomly.

### 3. RESULTS AND DISCUSSION

**McFarland's Principle.** A typical QSAR model considers a term for bioavailability and a term for the reactivity of the chemicals (McFarland's principle).<sup>25</sup> Generally, the bioavailability of a chemical to the organism is well described by the logarithm of the partition between octanol and water, i.e., LogP. Since octanol can represent the cell membrane, LogP indicates the ability of the chemical to permeate it and therefore to be available for interaction with the organism. This is especially true for aquatic toxicity where in the toxicological assay the target species is put into a solution with a given concentration of the chemical studied.

The other term is related to the chemical reactivity and HOMO (Highest Occupied Molecular Orbital) and LUMO (Lowest Unoccupied Molecular Orbital) are considered to be good representatives for it. These are called the frontier orbitals and govern the way the molecule interacts with other species.

Table 2. Summary View of the Results on the Data Set with 319 Descriptors

	train		test	
	$R^2$	rmse	$R^2$	rmse
MLR	1.00	0.000	0.00	10.526
PLS (2 components)	0.53	1.063	0.45	1.188
BPNN	1.00	0.000	0.00	2.888

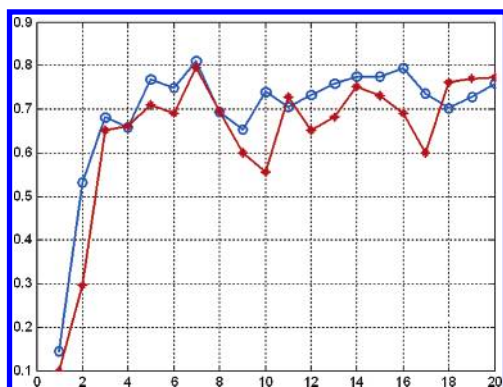
For this reason the first models were built up using only LogP, HOMO, and LogP, and LUMO descriptors.

The models in Table 1 are by far not satisfactory but provide some encouraging indications. As in previous publications,<sup>26–28</sup> the mechanism is confirmed for aquatic acute toxicity (coefficients for LogP are very similar in both the models and HOMO and LUMO have correctly opposite signs, since they describe opposite tendencies). Moreover, a narcotic baseline effect can be recognized.<sup>29</sup> The only chemical not conforming to the baseline effect is antimycin A (1397-94-0), which is an antibiotic with a peculiar mode of action and is likely to be an outlier.<sup>30</sup> Waes et al.<sup>31</sup> showed that the use of a more lipidlike system, i.e., log  $K_{DMPC}$ , can improve the quality of the model, nevertheless the low performances suggest that there are other mechanisms than narcosis, and the toxicological action of pesticides needs more parameters to be correctly described.

**Models on the Whole Data Set.** Different statistical techniques were used to extract information from the data set with 319 descriptors. Table 2 summarizes the results obtained with MLR, PLS, and BPNN.

A first general conclusion from the above is that the reliability of QSAR models has to be assessed on their predictive abilities rather than on the fitting properties that





**Figure 1.** Performance of the model in relation to the number of variables ( $x$ -axis). Blue circles ( $\circ$ ) are  $R^2_{\text{train}}$ ; red asterisks ( $*$ ) are  $R^2_{\text{test}}$ .

can be arbitrarily good. Considering the large difference between performances on the training set and the test set it can be stated that the problem is *ill-posed*,<sup>32</sup> and the selection of relevant variables is required. Moreover, nonlinear techniques, such as BPNN, seem to cope better with the complexity of the problem. For this reason a combination of genetic algorithm and counterpropagation neural network (GA/CPANN) was chosen to support our study.

**Selection of Descriptors, Development, and Interpretation of the Model.** A fundamental step in QSAR studies is interpretation of the model. It is good practice to allow a mechanistic and/or biological explanation of the statistical model. Clearly if there are only a few variables the model will be easier to interpret. On the other hand, reducing the number of variables decreases the model's ability to explain a complex phenomenon like toxicity. Figure 1 shows how the determination coefficient changes with respect to the number of variables for the training ( $R^2_{\text{train}}$ ) and test ( $R^2_{\text{test}}$ ) sets. The analysis was conducted with models using between 1 and 20 variables selected by GA/CPANN.

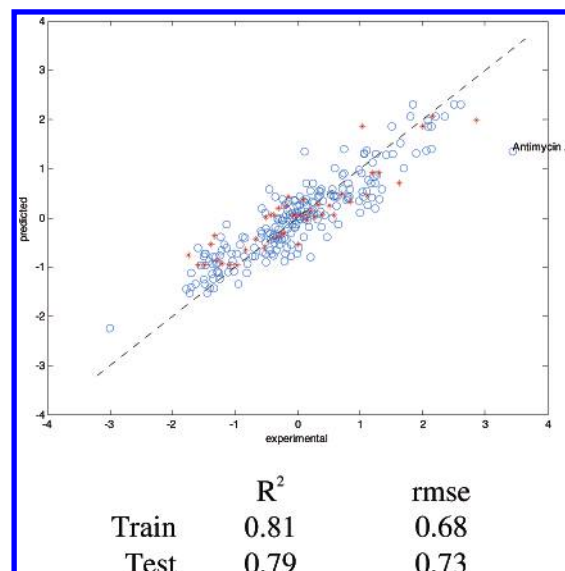
The model reaches a steady state after seven variables, and any new variable added is redundant. The small variability around recognizable trends is due to the intrinsic stochastic nature of the modeling technique. GA search for the "best" solution in the solution space, but since constraints are set, e.g. maximum number of generations or goal, the process may end in a local minimum.

Figure 2 shows the best model obtained by GA/CPANN using seven descriptors.

Again antimycin A showed to be an outliers. The descriptors selected are listed in Table 3.

The descriptors used in the best predictive model can be divided into two main categories: penetration/solubility descriptors, which reflect the compound's abilities to form noncovalent interactions with the environment, to dissolve and persist in water or a lipidic environment, or permeate the phase interfaces (i.e. LogP, hydrogen bonding descriptors, polarizability); and reactivity descriptors, which indicate the compound's abilities to interact with the surrounding molecules and form chemical bonds (i.e. orbital gap).

Another criterion for classifying descriptors is their dependence on the 3D structure. LogP and  $^3\chi_p$  are descriptors independent of the 3D conformation, while for other descriptors a 3D (optimized) conformation must be calculated. As in the OpenMolGRID system an automatic modeling procedure was applied to the structures, no conformational



**Figure 2.** GA/CPANN final model. Blue circles ( $\circ$ ) are chemicals in the train set; red asterisks ( $*$ ) are chemicals in the test set.

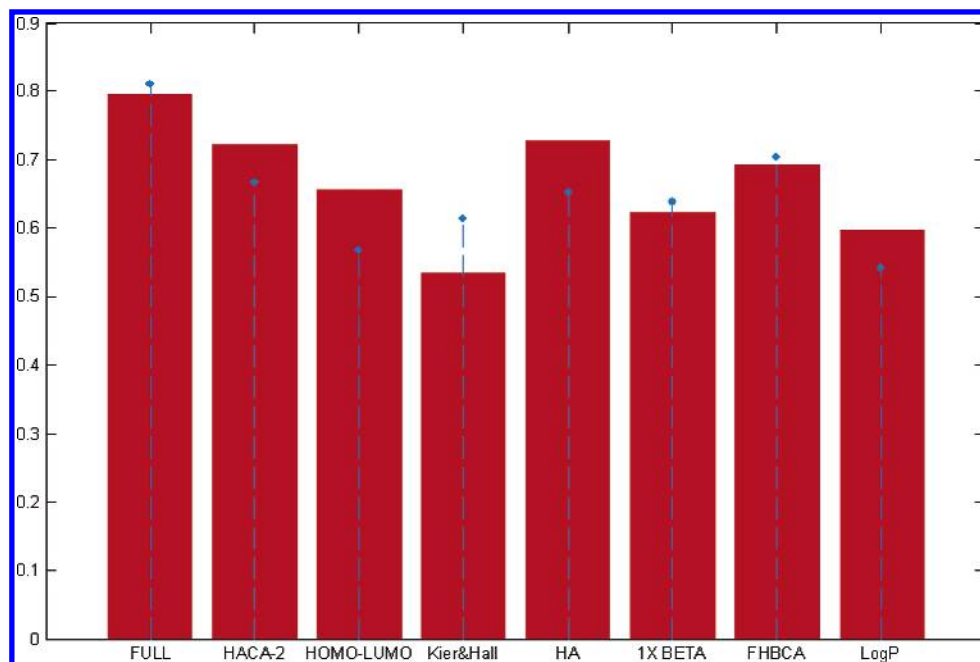
**Table 3.** Descriptors Selected by GA/CPANN

ID	descriptor
1	HACA-2 (MOPAC PC)
2	HOMO–LUMO energy gap
3	$^3\chi_p$
4	HA dependent HDSA-1 (Zefirov PC)
5	1XBETA polarizability (DIP)
6	FHBCA fractional HBSA (HBSA/TMSA) (MOPAC PC)
7	LogP (KowWin1)

search was done before optimization, and the resulting 3D conformations are probably the local minima rather than global ones. However, the promising results of the best predictive model presented here show that for a heterogeneous group of compounds "any reasonable" (i.e. optimized) conformation can be used to derive 3D descriptors and construct a predictive model. This is in clear contrast to the 3D approaches used in other areas of computer-aided molecular design (especially drug design or nanotechnologies), where the lowest energy 3D conformations for a series of conformationally (structurally) similar compounds are always sought. In any case, there are no evidences that the lowest energy conformation is actually related to the toxic mechanism. Although descriptors involved are not dramatically dependent from 3D conformations, some can encode spatial information.

Nonlinear models such as CPANN are usually more powerful than linear ones but are often considered "black boxes" because they do not formalize the relationship between variables and response in clear numbers or coefficients. This raises some doubts about their usage and reliability. Nevertheless, techniques to estimate the influence and weights of each variables to the model can be easily implemented. One of them is to substitute variables by a constant, for example its mean, and see how this affects the prediction. This is equivalent to neglect of the information encoded by them. Of course, the more a descriptor is relevant to the model the more the overall performance drops when it is substituted by a constant value. Figure 3 shows results obtained by the procedure on this model.

In general, the LogP (KowWin1) is a measure of the compound's lipophilicity. In aquatic species LogP describes



**Figure 3.** Descriptor analysis, showing the influence of setting variables at a constant value (mean) on the overall performances of the model. Blue stems are  $R^2_{\text{train}}$ ; red areas are  $R^2_{\text{test}}$ .

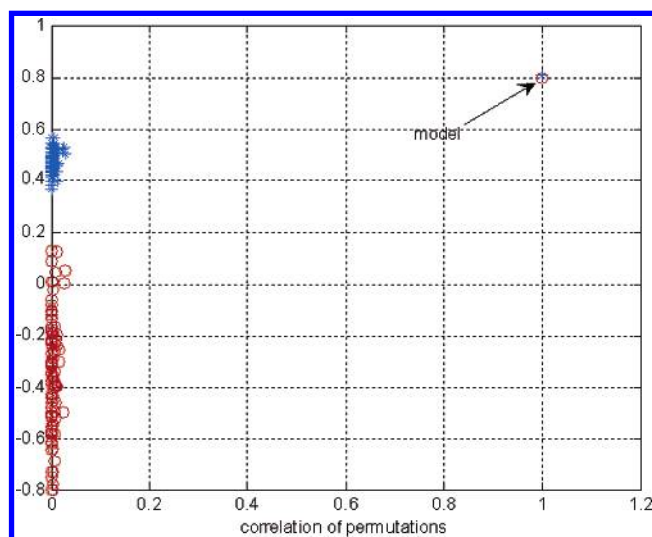
a compound's penetration and distribution in the organism. From the physical point of view LogP describes the entropic contributions which are important for solvation/desolvation. Numerous studies in aquatic toxicology have shown LogP to be an important descriptor frequently occurring in predictive QSAR models.

$^3\chi_p$  (i.e. path-3 Kier and Hall index) is a topological descriptor encoding the adjacency of branch points in the molecular skeleton.<sup>33</sup> Thus, in the resulting QSAR model it can be regarded as the descriptor of molecular structure in terms of connectivity. The Kier and Hall index is most important as the determination coefficients on the training set and on the test set decrease significantly if this descriptor is kept constant (the mean value) in the descriptor test (Figure 3).

The descriptors HACA-2 (MOPAC PC)—area-weighted surface charge of hydrogen bonding acceptor atoms, HA dependent HDSA-1 (Zefirov PC)—hydrogen bonding donor ability of the molecule, and FHBCA fractional HBSA (HBSA/TMSA) (MOPAC PC)—fractional hydrogen bonding surface area divided by total molecular surface area—all belong to the group of surface area descriptors related to hydrogen bond formation, intermolecular interactions, and compound solvation in the water environment. The apparent redundancy of hydrogen bonding-related descriptors in the model (three out of seven) is likely caused by the presence of the different subgroups of compounds in the database (e.g. sulfonamides, carbamates) and their importance for the toxicity of these groups of compounds.

The polarizability descriptor 1XBETA (BETA polarizability DIP) reflects the molecule's properties from the point of view of polarization induced by an external electric field and characterizes the molecule as an electron acceptor. This descriptor is very important for the model, as its replacement by the constant value (the mean) significantly decreases the test determination coefficient (Figure 3).

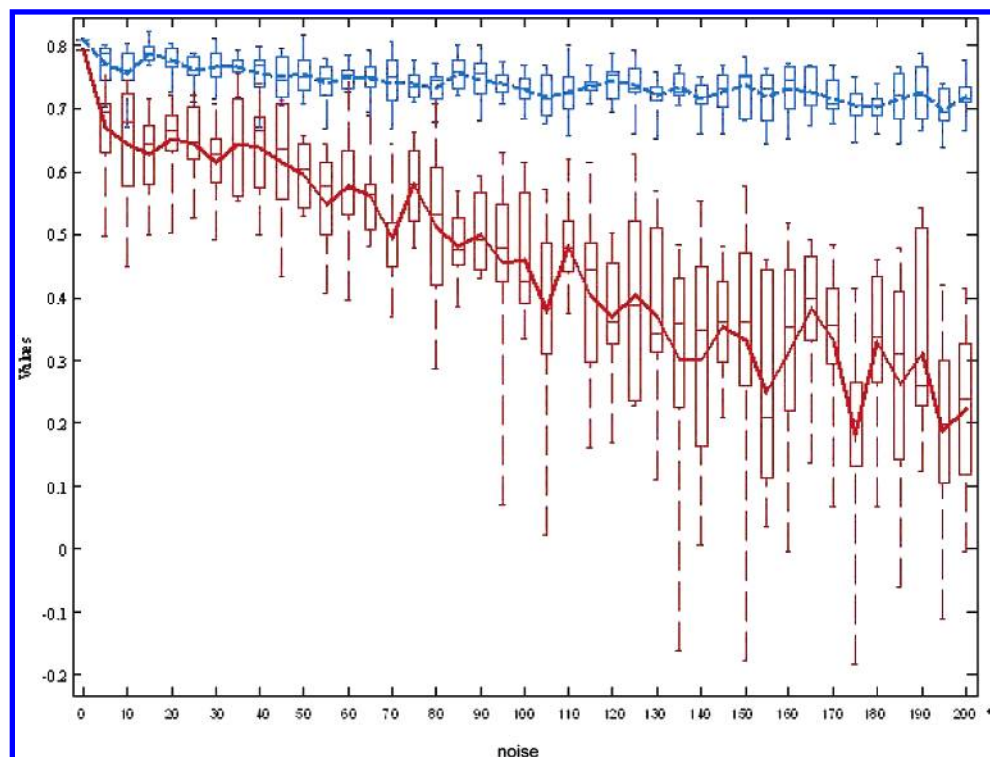
The descriptor *HOMO–LUMO energy gap* can be regarded as the descriptor of reactivity. The HOMO–LUMO



**Figure 4.** Response permutation testing: on the y-axis the performances of the model ( $R^2_{\text{train}}$ , as blue circles,  $R^2_{\text{test}}$  as red asterisks), and on the x-axis the correlation between original and scrambled response.

gap, i.e., the difference in energy between the highest occupied and lowest unoccupied molecular orbital, determines the compound's stability. The greater the difference, the lower the reactivity of the molecule. As it is the only "true" descriptor of the chemical reactivity in the model, it contributes substantially to the overall performance (Figure 4) in terms of both  $R^2_{\text{train}}$  and  $R^2_{\text{test}}$ .

**Additional Testing and Validation of the Best Predictive Model.** The model was subsequently validated using the response permutation test, also known as y-scrambling.<sup>34,35</sup> This procedure involves fitting several models, in our case 100, on the same dependent variables (X block) but on a permuted response. If a strong correlation remains between the descriptors selected and the randomized response, then the significance of the proposed QSAR model is regarded as suspect.



**Figure 5.** Sensitivity test. Ten models are fitted for each level of noise. Boxes have lines showing the lower quartile, median, and upper quartile of each level of noise. The whiskers extend from each end of the box to show the extent of the rest of the data.  $R^2_{\text{train}}$  is shown in blue (hatched),  $R^2_{\text{test}}$  in red. Marked lines show the means for each level of noise.

In our case (Figure 4), the model performs much better than any of the permuted models. Although models on random numbers show some recall ability, i.e., acceptable  $R^2_{\text{train}}$ , they fail to predict the external test. This is clear proof that the model is not affected by any chance correlation, i.e. the probability to obtain a similar or better model using random numbers is null, and it is likely to depict a true relationship.

A further test was done to assess the reliability of the model. A sound model should be stable and not too sensitive to noise. To simulate the influence of noisy data on the performance of the model, we added some randomness to the X block. Given the  $r$ -by- $c$  matrix of the X block, where  $r$  is the number of objects (chemicals) and  $c$  the number of descriptors, for each  $c$ th column  $r$  uniformly distributed random numbers in the interval (0,1) are calculated, scaled by a given percentage  $n$  of the standard deviation of the  $c$ th descriptor and added as noise to the column. Thus we have a new  $r$ -by- $c$  matrix which is the original X block plus a given noise  $n$ . This data set is then used to fit a model to predict the original response. For each noise  $n$  50 runs were done. Results are summarized in Figure 5.

Figure 5 nicely illustrates a behavior common to the majority of neural networks. Observing  $R^2_{\text{train}}$  (blue lines) the model “learns” the noise as well as the “true” data. This is known as overfitting and occurs when a learning algorithm is allowed adapting too well the training set, using for example too many variables and/or training epochs. Hawkins defines overfitting as “the use of models or procedures that violate the principle of parsimony, ... or Occam’s Razor” and gives an exhaustive description of the problem.<sup>22</sup> The reliability of QSAR models must therefore be assessed on their predictive ability rather than on the fitting properties which may arbitrarily be good.

**Table 4.** Validation of the Descriptors Used for the Acute Aquatic Toxicity Model

	DASTD	DVR	DVP%
HACA-2 (MOPAC PC)	0.276	10.277	2.7
HOMO–LUMO energy gap	0.295	11.342	2.6
$^3\chi_p$	0.000	15.181	0.0
HA dependent HDSA-1 (Zefirov PC)	4.504	186.560	2.4
1X BETA polarizability (DIP)	131.080	6232.900	2.1
FHBCA fractional HBSA	0.015	0.379	3.9
(HBSA/TMSA) (MOPAC PC)			
LogP (KowWin1)	0.000	20.57	0.0

In line with these general comments, we focused on the predictive power of the model, i.e.,  $R^2_{\text{test}}$  (red lines). This test itself of course does not ensure that the model reflects a true relationship, but points to a reassuring behavior, stability: the model performances smoothly worsen as the noise increases but do not present any chaotic phenomenon, or in other words there is no sensitive dependence on initial conditions. Moreover, during this test “new” objects are generated, that far from being real chemicals are at least acceptable in that their descriptor values are not too distant from real values (the perturbation added is a fraction of the standard deviation). Again, the model generates reasonable prediction in relation to possible inputs.

Molecular descriptors (MD) are the result of mathematical operations which transform the chemical information encoded within a symbolic representation of a molecule. Unfortunately, such a numerical representation is not unique. Indeed, each descriptor is expected to show a variability which depends strongly on the level of chemical theory behind it. For example, 2D constitutional descriptors will not change with molecular conformation. Regardless of the computational chemistry method used, the values of these descriptors are expected to match each other perfectly. On



**Table 5.** Boundaries of Property and Relevant Descriptors for the Three Sets

	train		test	
	min	max	min	max
Log <sub>10</sub> (1/LC <sub>50</sub> )	-2.335	7.7393	-0.3567	6.8438
HACA-2 (MOPAC PC)	0	10.915	0	6.8352
HOMO-LUMO energy gap	2.935	14.7	6.718	12.101
$\chi_p^v$	0	15.181	0.27386	11.306
HA dependent HDSA-1 (Zefirov PC)	0	174.15	0	129.3
BETA polarizability (DIP)	-993.29	1041.2	-681.23	366.43
FHBCA fractional HBSA (HBSA/TMSA) (MOPAC PC)	0	0.38393	0	0.22627
LogP (KowWin1)	-5.92	9.82	-1.1	8.39

the other hand, 3D descriptors, especially quantum-mechanical ones, are much more sensitive than any other descriptors with respect to molecular structure. In fact, the use of different optimization procedures leads to different 3D geometries thus to different values of 3D molecular descriptors. The key point of the current investigation is not to quantify the MD variability exactly but to determine to what extent these values are comparable to each other. Having comparable MD values means having a QSAR model that is not dramatically dependent on the exactness of the 3D chemical structure.

To make this analysis, three sets of descriptors using respectively MNDO, PM3, and AM1 methods<sup>36</sup> have been generated by the descriptor calculation workflow described earlier and then analyzed using the following criteria:

1. Descriptor Average Standard Deviation (DASTD), defined as the mean standard deviation of each value of the *j*th descriptor:  $DASTD_j = \sum std(D_{i,j})/n$ .

2. Descriptor Variability Range (DVR), defined as the difference between the maximum and the minimum value of the *j*th descriptor:  $DVR_j = \text{Max}(D_j) - \text{Min}(D_j)$ .

3. Descriptor Variability Percentage (DVP%), defined as  $DVP\%_j = DASTD_j/DVR_j \cdot 100$ . This parameter indicates the average variability within the maximum range of possible values it assumes. DVP% does not depend on the absolute value of a single descriptor, providing a concrete mean to compare the variability of diverse descriptors.

Having *n* compounds and *m* descriptors,  $D_{i,j}$  are the values that the *j*th descriptor has for the *i*th structure according to the three different parametrizations, and  $D_j$  are all the values of the *j*th descriptor.

The results above (Table 4) indicate an excellent consistency of the descriptors relevant to the QSAR model proposed, among different 3D geometries. This fact allows for a wider and simpler applicability of the model.

The definition of applicability domain of any QSAR is still an open issue, because it raises doubts about the validity of interpolation and/or extrapolation in multidimensional spaces.<sup>37,38</sup> Anyway boundaries (see Table 5) are usually useful in order to assess the chemical space of QSARs. Although training and test sets have generally similar boundaries, some gaps are indeed present, due to the fact the test set was randomly selected. The large difference between toxicological minimum values is caused by the presence in the training set of 2-propanol (67-63-0) that is largely less toxic than any other pesticide in the data set. The real applicability domain of the model should then be considered the chemical space located within the test set.

#### 4. CONCLUSIONS

The main outcome of this study is the development of a predictive model for aquatic acute toxicity. The rigorous

procedure adopted to test the QSAR model ensures its applicability and reliability in predicting toxicity of new (not yet tested) pesticides, making it particularly suitable for regulatory purposes, which is the main goal of the DEMETRA project. The mechanism emerging from analysis of the model and its descriptors is consistent with McFarland's principle for biological activity, i.e., the activity (toxicity) of a given compound is the sum of the compound's abilities to penetrate (lipophilicity, hydrophilicity) and interact with biological structures (reactivity).

Generally classical linear methods can provide useful preliminary information but cannot solve complex QSAR problems. They may work on a local model but not for a structurally diverse database like the trout LC<sub>50</sub> studied in this work. Nonlinear methods such as neural networks cope with the complexity of the problem but they dramatically suffer from overfitting, so a features selection is essential to reduce intrinsic variability and improve the generalizability of the model. Among the different techniques tested, the GA/CPANN combination proved suitable for the development of ecotoxicological QSARs, because it can extract useful information hidden in the numbers and it is flexible enough to detect the nonlinear relationships between molecular descriptors and biological activity.

#### ACKNOWLEDGMENT

This work is partially funded by EU projects DEMETRA (EU FP5 HPRN-CT-1999-00015) and OpenMolGRID (EU FP5 IST-2001-37238). The collaboration of project partners is gratefully acknowledged. Paolo Mazzatorta thanks Prof. Mark T. D. Cronin for his careful, helpful supervision.

**Supporting Information Available:** The full list of chemicals involved in this work. This material is available free of charge via the Internet at <http://pubs.acs.org>.

#### REFERENCES AND NOTES

- (1) Pimtel, D. Amounts of pesticides reaching target pests: environmental impacts and ethics. *J. Agric. Environ. Ethics* **1995**, *8*, 47-84.
- (2) WHO-UNEP. *Public Health impact of pesticides used in agriculture*; World Health Organization-United Nations Environment Programme, Geneva, Switzerland, 1989.
- (3) OECD. OECD Guidelines for Testing Chemicals, METHOD 203, Fish, Acute Toxicity Test, Paris, 1984.
- (4) OPP 72-1. Acute Toxicity Test for Freshwater Fish (Pesticide Assessment Guidelines, Subvision E-Hazard Evaluation; Wildlife and Aquatic Organisms), EPA report 540/09-82-024, 1982.
- (5) OPPTS 850.1075. Ecological Effects Test Guidelines, Fish Acute Toxicity Test, Freshwater and Marine.
- (6) Commission of the European Communities. *White Paper on Strategy for a Future Chemicals Policy*; Bruxelles, February 27, 2001, COM 88 final.
- (7) Omen, G. S. Assessing the Risk Assessment Paradigm. *Toxicology* **1995**, *102*, 23-28.

- (8) Walker, J. D. QSARs for Toxicity Screening: Current Practices. In *QSARs for Pollution Prevention, Toxicity Screening, Risk Assessment, and Web Applications*; Walker, D., Ed.; Published by SETAC: 2003.
- (9) Schultz, T. W.; Cronin, M. T. D. Essential and desirable characteristics of ecotoxicity quantitative structure–activity relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 599–607.
- (10) Roncaglioni, A.; Benfenati, E.; Boriani, E.; Clook, M. A Protocol to Select High Quality data sets of Ecotoxicity Values for Pesticides. *J. Environ. Sci. Health B* **2004**, *39*, 641–650.
- (11) Mazzatorta, P.; Benfenati, E.; Schuller, B.; Romberg, M.; McCourt, D.; Dubitzky, W.; Sild, S.; Karelson, M.; Papp, A.; Bágyi, I.; Darvas, F. OpenMolGRID: Molecular Science and Engineering in a Grid Context. In *Proceedings of PDPTA 2004, The 2004 International Conference on Parallel and Distributed Processing Techniques and Applications*; June 21–24 2004, Las Vegas, Nevada, U.S.A., 2004.
- (12) <http://www.openmolgrid.org/>.
- (13) Herman W. Partial Least Squares. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; Wiley: New York, 1985; Vol. 6, pp 581–591.
- (14) Funahashi, K. I. On the approximate realization of continuous mappings by neural networks. *Neural Networks* **1989**, *2*, 183–192.
- (15) <http://www.mathworks.com/access/helpdesk/help/toolbox/nnet/>.
- (16) Holland, J. *Adaptation in Natural and Artificial Systems*; The University of Michigan Press: Ann Arbor, MI, 1975.
- (17) Zupan, J.; Novic, M.; Gasteiger, J. Neural networks with counter-propagation learning strategy used for modelling. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 2, 175–187.
- (18) Mazzatorta, P.; Vračko, M.; Jezierska, A.; Benfenati, E. Modeling toxicity by using supervised Kohonen neural network. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 485–492.
- (19) Mazzatorta, P.; Vračko, M.; Benfenati, E. ANVAS: Artificial Neural Variables Adaptation System for descriptor selection. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 335–346.
- (20) Topliss, J. G.; Costello, R. J. Chance Correlation in Structure–Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972**, *15*, 1066–1068.
- (21) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure–Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.
- (22) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
- (23) Mazzatorta, P.; Benfenati, E.; Neagu, C. D.; Gini G. The importance of scaling in data mining for toxicity prediction. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1250–1255.
- (24) Golbraikh, A. Molecular data set Diversity Indices and Their Application to Comparison of Chemical Databases and QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 414–425.
- (25) McFarland, J. W. On the Parabolic Relationship Between Drug Potency and Hydrophobicity. *J. Med. Chem.* **1970**, *13*, 1192–1196.
- (26) Könemann, H. Quantitative structure–activity relationships in fish toxicity studies Part 1: Relationship for 50 industrial pollutants. *Toxicology* **1981**, *19*, 3, 209–221.
- (27) Cronin, M. T. D.; Schultz, T. W. Validation of *Vibrio fischeri* acute toxicity data: mechanism of action-based QSARs for nonpolar narcotics and polar narcotic phenols. *Sci. Total Environ.* **1997**, *204*, 1, 75–88.
- (28) Ramos, E. U.; Vermeer, C.; Vaes, W. H. J.; Hermens, J. L. M. Acute toxicity of polar narcotics to three aquatic species (*Daphnia magna*, *poecilia reticulata* and *Lymnaea stagnalis*) and its relation to hydrophobicity. *Chemosphere* **1998**, *37*, 4, 633–650.
- (29) Öberg, T. A QSAR for Baseline Toxicity: Validation, Domain of Application, and Prediction. *Chem. Res. Toxicol.* **2004**, *17*, 12, 1630–1637.
- (30) Lipnick, R. L. Outliers: their origin and use in the classification of molecular mechanisms of toxicity. *Sci. Total Environ.* **1991**, *109–110*, 131–153.
- (31) Waes, W. H. J.; Ramos, E. U.; Verhaar, H. J. M.; Hermens, J. L. M. Acute toxicity of nonpolar versus polar narcotics: is there a difference? *Environ. Toxicol. Chem.* **1998**, *17*, 1380–1384.
- (32) Hadamard, J. *Lectures on the Cauchy Problem in Linear Partial Differential Equations*; Yale University Press: 1923.
- (33) Hall, L. H.; Kier, L. B. Molecular connectivity: intermolecular accessibility and encounter simulation. *J. Mol. Graphics Modell.* **2001**, *20*, 76–83.
- (34) Van der Voet, H. Comparing the predictive accuracy of models using a simple randomization test. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 313–323.
- (35) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis – Principles and Applications*; Umetrics AB: Umea, Sweden, 2001.
- (36) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*; ISBN: 0-471-48552-7, 562 pages.
- (37) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. Review of methods for QSAR applicability domain estimation by the training set. *ATLA* **2005**, in press.
- (38) Eriksson, L.; Jaworska, J.; Worth, A.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression- Based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1351–1375.

CI050247L