

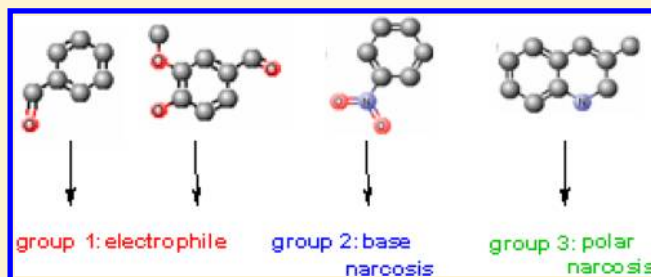
# Determination of Toxicant Mode of Action by Augmented Top Priority Fragment Class

Mosé Casalegno<sup>‡</sup> and Guido Sello<sup>\*,§</sup>

<sup>‡</sup>Department of Chemistry, Materials, and Chemical Engineering, "Giulio Natta", Via Mancinelli 7, I-20131 Milano, Italy

<sup>§</sup>Dipartimento di Chimica, Università degli Studi di Milano, via Golgi 19, I-20133 Milano, Italy

**ABSTRACT:** Theoretical models can be an efficient tool to assess compound toxicity as an alternative to experimental determinations. Their application must follow some requirements that include the possibility of understanding the rationale that supports the prediction; here, the determination of the mode of action (MOA) is important. A combination of similarity and reactivity analysis has been applied to group chemical compounds with the aim at selecting groups that share structure and electronic state. The model is not based on experimental data but only on structural features. The result is a number of groups that contains similar compounds with similar reactivity and, possibly, similar MOA. The comparison of these groups to the experimentally determined MOAs available for the EPAFHAM database permits the discussion of the validity of both the model and the experimental data.



## 1. INTRODUCTION

The REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) regulation, established in 2007, calls for the classification of the toxic attributes of all the substances produced or imported in more than 1 ton per year, completed by data concerning their effects on the environment and the human health.<sup>1</sup> In this regard, QSAR (Quantitative Structure Activity Relationship) studies can be used, in the view that similar compounds should have the same toxic effect and/or mode of action (MOA). The MOA term has many meanings;<sup>2</sup> however, two of them are of special interest: the reference to a specific physiological effect and the proposition of a reaction mechanism at molecular level. Even if we can think that these two roles are strictly connected, the available data show that the variability of the molecular characteristics is much richer than the experimental effects it produces. As a consequence, the connection between the molecular moieties and their molecular action is difficult to be determined. An essential common point of MOA models is the WOE (weight of evidence), mainly based on Hill<sup>3</sup> suggestions, because they can be used as a list of criteria; e.g.: dose–response relation, experimental proofs, biological plausibility, data consistency, and reproducibility.

MOA classification has been also applied to the environmental risk assessment.<sup>4,5</sup> Here, the data extrapolation from species to species is more common, because the analysis should be more general. The data used to classify the compounds often come from acute toxicity tests, ACR (acute to chronic ratios), used to calculate the LCS0 that are often the only available experimental data. Many toxic effects act on the fundamental cell components that are common to several organisms;

however, there are exceptions, and it is important to link the biological target to the effect in order to get reliable models.

The importance of the knowledge of the mechanism of action is clear: if the biological target is known as is the action of the compound, it will be possible to assign the right descriptors for QSAR, passing from the structural similarity to action similarity to toxicity similarity. In addition, this approach can help in making explicit the confidence level, providing a more transparent decision path.

Concerning aquatic organisms, more than 50% of the toxicants have nonpolar nonspecific toxic effects;<sup>6</sup> the rest have a greater toxicity in the following MOAs: polar nonspecific, uncouplers of oxidative phosphorylation, inhibitors of acetylcholinesterase, inhibitors of respiration, thiol-alkylating agents, reactives. It is complicated to develop QSAR models that use the same descriptors in all MOA cases. Thus, the alternative is to have a first model that classifies the compounds in MOA groups and a set of QSAR models that are specific for each group. This can be realized by developing a procedure that generates clusters of compounds that are supposed to share the same MOA. Most of the studies are still using as reference the EPAFHAM<sup>7</sup> data set that lists an experimental MOA for many compounds. Two recent examples<sup>8,9</sup> implement a similar approach: a set of molecular descriptors is selected; a training set of compounds is classified using the descriptor set as input variables and the MOA as output variable; the result is checked against the EPAFHAM data set. All steps are repeated until a good agreement between the calculated classes and the experimental MOAs is reached. Then, several tests are

Received: February 28, 2013

performed to verify the result quality. The two approaches use different statistic functions and very different molecular descriptors, but their fundamentals are very similar.

Continuing our studies in the field of toxicity prediction, we took into consideration the possibility of compound classification. Our approach would like to go a step further in the analysis of the problem of compound classification, in the attempt of making the classification independent from the experimental data, that will be used only as a subsequent control. In a previous paper<sup>10</sup> we used a similar approach in the definition of a domain of applicability that is completely unconnected to the QSAR model. The motivation has two supports: first, a general approach can be applied to many cases; second, the experimental data are sometimes not sufficiently reliable.

In considering MOA for aquatic toxicity, we should consider the intrinsic differences between the large class of narcotics and the other mechanisms. Narcosis is an unspecific effect whose corresponding biological target is unclear. In contrast, the other mechanisms can be seen as triggered by a specific compound moiety. This difference is highly relevant; in fact, at the molecular level, narcosis is an effect that must be always present; consequently, narcosis is sometimes referred to as base toxicity. When a special mechanism is also present the narcosis effect is more or less hidden. The corollary is that we should first locate the special mechanisms and then assign everything else to narcosis classes.

In past works<sup>11,12</sup> we have used a special tool to assign compounds to classes; i.e. we tagged each compound with several tags, and then we organize the tags obtaining an ordered list of memberships that could be used in different topics. We made the assumption that all compounds that share the same tag have a description similar enough to show the same behavior. The tag has also the function of directing the compound comparison during the classification. In this work we aim at extending the tag approach to a more articulated scheme that can be appropriate for compound classification in diverse MOAs. It is important to note that the procedure is not based on experimental data, i.e. it is a completely blind procedure; thus, it does not need training test sets.

## 2. EXPERIMENTAL PROCEDURE

**2.1. Data.** The data are derived from two main sources: the EPAFHAM and the Demetra databases. Concerning the EPAFHAM database some more data elaboration are derived from the literature.<sup>6,13–21</sup>

**2.2. Method.** **2.2.1. Overview.** Before describing the method used in the present approach we would like to provide a qualitative view of its fundamentals. In the previous paper concerning the determination of the applicability domain we developed a method for outlier detection that begins with the characterization of the molecular structures on the basis of fragments.<sup>10</sup> Using fragments we generated a chemical space where the discrete nature of such descriptors was exploited by the realization of a cluster-based approach that permits the outlier detection in an indirect way. The basic idea was to map the fragment-based representation onto a cluster-based one. Clusters were groups of compounds sharing a common fragment. Thus, each compound belonged to as many clusters as the number of its constituent fragments. This introduced a descriptive framework where compounds were described by means of membership in clusters (also referred to as sets), and a function called affinity quantified the tendency of a

compound to belong to a cluster. The problem we wanted to solve was the quantification of the contribution of each cluster to the compound description. At the end of the calculation we had sets containing some compounds; a compound could be present in more than one set.

Because the sets are formed starting from fragments they have not any special reactivity meaning, but another approach<sup>22</sup> we developed several years ago was based on molecular electronic properties. In that paper we introduced a new definition of functional groups and a method for their identification. The basis of the definition was the introduction of a measure of the importance of each non-hydrogen atom in a molecule. Then, using a walk through the molecule we located the functional groups selecting the most important atoms and appending to them all neighboring less important atoms. The hypothesis was that the dominant atom is electronically connected to the atoms it can influence; i.e., we defined a functional group as a set of interacting fellows. It is clear that this definition is directly connected to reactivity. The use of these functional groups is, in contrast, inappropriate to cluster toxic compounds because their description is too detailed.

The objective of this work is to get the best from these two approaches and to apply this method to classify compounds; if the model is working the formed sets should contain compounds with similar structure and reactivity, and they can be used to assign a common MOA. A brief description of these methods is worth, even if a more exhaustive description can be found in the original papers.<sup>10,22</sup>

**2.2.2. The Determination of Structural Fragments (SF) via Recursive Clustering.** The determination of the SFs is performed using a recursive clustering approach. Hereafter, we briefly resume the mathematical framework of this method.

We consider a chemical structure database containing  $N$  compounds, characterized by a set of  $M$  molecular substructures, referred to as fragments. For each molecule, the binary occurrence of each fragment is computed, and an occupancy matrix  $O$ , consisting of  $N \times M$  elements, is filled as follows:

$$O(i, j) = \begin{cases} 1, & \text{if the } j\text{-th fragment occurs in the } i\text{-th molecule one or more} \\ 0, & \text{if the } j\text{-th fragment is absent in the } i\text{-th molecule} \end{cases} \quad (1)$$

We also define a pairwise similarity matrix  $S$ , consisting of  $N \times N$  elements, whose generic element  $S(i, ii)$  quantifies the structural similarity between the  $i$ -th and  $ii$ -th molecules, defined as follows

$$S(i, ii) = \begin{cases} R(i, ii), & \text{if } i \neq ii \\ 0, & \text{if } i = ii \end{cases} \quad (2)$$

where  $R(i, ii)$  is the numerical value assumed by the pairwise similarity function for the  $i$ -th and the  $ii$ -th molecules, a real number in the interval  $[0, 1]$ . The pairwise similarity matrix serves to drive the recursive evaluation of the clustering process.

At this point, we count the total number of molecules where the  $j$ -th fragment occurs,  $n(j)$ , defined by the following equation:

$$n(j) = \sum_{i=1, N} O(i, j) \quad (3)$$

Then, we associate to each fragment a set  $C(j)$ , defined as the collection of  $n(j)$  molecules where the  $j$ -th fragment occurs, i.e. the cluster of compounds that share the  $j$ -th fragment. The sets

associated with singly occurring fragments (SOF) contain only one compound and will not be further considered. The definition of  $C(j)$  tells us that each molecule belongs to all clusters that can be associated with its constituent fragments. The number of clusters the  $i$ -th molecule belongs to, after removal of SOFs, can be computed in the following way:

$$m(i) = \sum_{j=1,M} O(i, j), \text{ for } n(j) > 1 \quad (4)$$

The memberships of each molecule across different clusters are set equal to  $1/m(i)$ , which is the fraction associated with the number of clusters accessible to the  $i$ -th compound. The corresponding weights are then equal to unity. This configuration gives all  $m(i)$  sets the same importance in describing a molecule, regardless of their composition, but sets containing molecules structurally similar to the queried compound should contribute more than others to the molecular description. At this point, we need to build a function to evaluate the descriptive ability of each set. We are particularly interested in determining the degree of structural similarity between the generic  $i$ -th molecule and those belonging to the  $m(i)$  sets associated with it. One simple possibility is to compute the weighted average similarity, hereafter called affinity

$$A(i, j) = \frac{(\sum_{ii \in C(j)} S(i, ii) \cdot W(ii, j))}{n(j)-1} \quad (5)$$

where

$$W(ii, j) = \begin{cases} O(ii, j), & \text{if } n(j) > 1 \\ 0, & \text{if } n(j) = 1 \end{cases}$$

Here,  $A(i, j)$  is the affinity between the  $i$ -th molecule and the set  $C(j)$ . We choose the term “affinity” to avoid confusion with the term “similarity”, which is used in the contest of pairwise comparisons. The new distribution, with a new membership matrix,  $P$ , describes each molecule more accurately within the clustering scheme defined by the used fragments. At the end of the model application a molecule belongs only to those sets that contain other similar molecules. Each set is tagged by the fragment used to form the set.

In this work, we shall use Atomic Centered Units (ACUs) as molecular fragments. This choice is primarily motivated by the need to deal with databases differing widely in size and chemical composition. A description of these fragments and of their generation through a molecular breakdown process has already been provided.<sup>23</sup> ACUs are small substructures, ranging from two to five atoms, made up by collecting a central parent atom and its closest neighbors. The small size and high statistical occurrence make them ideal candidates for our purposes.

The immediate availability of ACUs also suggested a simple functional form for the similarity function. To compute the pairwise similarity,  $S(i, ii)$ , we applied the well-known Tanimoto similarity formula<sup>24</sup> to the values from the occupancy matrix

$$S(i, ii) = \frac{C(i, ii)}{U(i) + U(ii) - C(i, ii)} \quad (6)$$

where

$$C(i, ii) = \sum_{j=1,M} O(i, j) \cdot O(ii, j) \quad (7)$$

$$U(i) = \sum_{j=1,M} O(i, j) \cdot O(i, j) \quad (8)$$

In the sums reported above, all fragments, including singly occurring ones, were explicitly considered.

**2.2.3. The Definition of Functional Groups (FG) via the Molecular Electronic Energy.** In this approach, a molecule is characterized by its electronic energy.<sup>22</sup> Its calculation can be performed using the well-known relation between electronic energy and chemical potential, as reported in eq 9

$$\mu = -X = (\delta E / \delta N)_Z \quad (9)$$

where  $\mu$  is the chemical potential,  $X$  is the atomic electronegativity,  $E$  is the electronic energy,  $N$  is the atom electronic population, and  $Z$  is the atom core potential that is considered constant in the derivative. The procedure for the calculation of the atomic energy is reported in the Appendix. The method permits the calculation of the electronic energy for each atom. In agreement with a much more complex approach, mainly due to Bader,<sup>25</sup> the calculation of the molecular energy can be made by summing the atomic contributions (eq 10)

$$E = \sum E_i \quad (10)$$

where the sum is over all the atoms in the molecule, and  $E_i$  is the energy of atom  $i$ .

The electronic energy of a molecule depends on the component atoms and their interactions. In fact, if an atom has significant interactions it stabilizes the molecule. In the molecular orbital sense, the combination of the atomic orbitals gives stable molecular orbitals if the attractive energies (core–electron) of the atoms are greater than the repulsive energies (both core–core and electron–electron), i.e. if the interaction of the electrons of one atom with the other nuclei stabilizes the molecule more than the interactions between the nuclei and/or between the electrons destabilize it. If the energetic contribution of an atom pair is high, i.e. if it highly stabilizes the molecule, we can say that the two atoms have a “strong” interaction. Atoms that have strong interactions are qualified to be FG components. The “importance” of an atom is, in this context, a quantity; it is thus possible to find the most “important” atoms and to call them “central” atoms. An FG is thus composed of a set of connected atoms that are considered sufficiently “important”; we only need the rule to distribute the atoms in different FGs. If we consider three connected atoms  $A-B-C$ , which respectively have  $E_A$ ,  $E_B$ , and  $E_C$  energy contributions, we can have four different situations:

(1)  $E_A > E_B > E_C$ , the importance is decreasing; all the atoms belong to the same FG.

(2)  $E_A > E_B < E_C$ ;  $E_A > E_C$  the importance is initially decreasing and then increasing; atom A is more important than atom C, A and B belong to one FG, and C belongs to another FG.

(3)  $E_A > E_B < E_C$ ;  $E_A < E_C$  the importance is initially decreasing and then increasing; atom C is more important than atom A, A belongs to one FG, and B and C belong to another FG.

(4)  $E_B$  is too low; therefore, B is not an FG atom, and A and C belong to two different FGs.

The procedure that actually searches and finds FGs is a straightforward application of the principles outlined in the

previous section. Its main activity is the calculation of the molecular energies for the given molecule (T) and for all the molecules obtained from T by isolating in turn one non-hydrogen atom from the others; the isolation of each atom is accomplished by cutting all the bonds connecting it to its neighbors, thus eliminating all the corresponding interactions. The comparison of the relative “importance” of each atom requires a set of homogeneous values. This set is composed of the (de)stabilization energies of all the molecules obtained from T. The energies are calculated by taking the difference between the electronic energy of each molecule containing an isolated atom and that of the reference T.

#### 2.2.4. Molecular Sets Formation, Merging, and Selection.

After using both the recursive clustering procedure and the functional group identification we have available two data categories: a number of clusters containing molecules that have the tag in common and that are similar enough, the tag is the SF, i.e. a selected ACU; the molecule representation as FGs, these are formed by a variable number of atoms grouped by their electronic interactions. We should add that a molecule can be part of more than one cluster. The first step is to relate the SFs to the FGs; this operation, hereafter called SF-FG mapping, can be easily accomplished by cross checking the atoms of the SF and of the FG: all the atoms of the SF should belong to one FG. If this is not the case, then the SF-FG mapping fails, and the current molecule is eliminated from the cluster. At the end of this step we will have a one-to-one connection between each SF and one specific FG for each molecule of the SF sets. In the Results section we shall provide one example of SF-FG mapping.

Because the SFs are rigidly defined, it can happen that some SF pair is very similar; the second step is therefore the analysis of the SFs to find indications of possible merging. This action, hereafter called cluster merging, will reduce the total number of SFs. To perform the action we again use the SF-FG relation. From a toxicological viewpoint, the descriptor associated with the FG is more significant than the descriptor associated with the SF; in the last case, the descriptors are classical bond-atom types that are assembled by exhaustive fragmentation of the molecular structure; in contrast, in the former case, the descriptors are determined by a physical molecular property and their composition is influenced by the molecular environment, i.e. they can be interpreted as an indication of molecular reactivity. The SFs that have corresponding atoms in the FG of similar weight, or that are a complete subset of the FG, can be merged and considered a single tag. Consequently, the molecules of the merged SFs are now part of the same cluster.

We can perform one more action on our data: we need to assign a molecule to a single cluster (cluster selection), at least as a starting point. For all those compounds that are still part of many sets, we should find a reason to select one set only. Because we are using the weights of the atoms in the FGs as a measure of reactivity, we are going to select that SF that contains the most important atoms. At the end of this step, we have a number of sets containing each molecule only once.

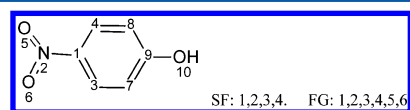
The last step is of some concern. As mentioned above, FGs are not of fixed length because they group all the atoms that are energetically interacting. The question is the following: is it possible that FGs of different lengths have different reactivity. To be sure, we check inside each group if there are molecules that are grouped using FGs of different lengths, and, in case, we subdivide these molecules. It can seem that we are canceling the

action we carried out during the merging phase; however, this is not the case. Again, in the Results section we will better explain the action by examples.

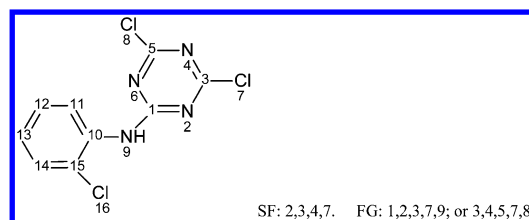
A brief summary of the overall strategy can help the assessment of our strategy. Essentially there are four possible actions that operate on the available clusters: 1) the SF-FG mapping; 2) the cluster merging; 3) the cluster selection; and 4) the cluster splitting. Action 1 is applied to all SFs and action 3 is applied to all compounds; in contrast, actions 2 and 4 are applied only to the cases that comply with the corresponding rules. The SF-FG mapping is the validation of the SF by the electronic properties of the molecule: it represents a first step toward a physical interpretation of the fragments. The cluster selection is only an attempt to fix a hierarchy in the cluster set. The cluster merging and splitting are two actions that intend to transform the clusters based on structure similarity into clusters that are also validated by electronic properties.

### 3. RESULTS

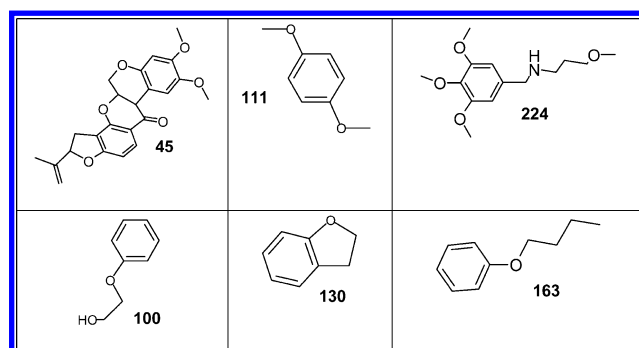
In this work, two databases were investigated. The first is the well-known and perused EPAFHAM database (hereafter



**Figure 1.** SF-FG mapping example. SF: atoms part of the structural fragment; FG: atoms part of the functional group.



**Figure 2.** SF-FG mapping example. SF: atoms part of the structural fragment; FG: atoms part of the functional group. No FG contains all SF atoms.

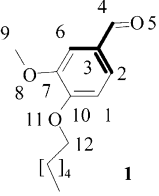
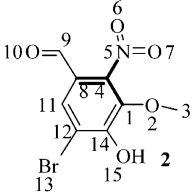
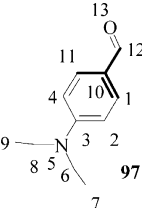


**Figure 3.** Cluster merging example. The molecules of clusters 7 and 9 are merged.

identified by the Duluth name); the second is a pesticide database, Demetra, also developed by the EPA organization. Duluth contains 617 compounds collected for their activity studies on fathead minnow fish. The database contains many data, among which there is the compound MOA; for 154 compounds the MOA is not determined due to insufficient or conflicting data. A confidence level is also assigned to each



Table 1. Example of Cluster Selection

Compound	Cluster number, SF atom, atom energy, and mean energy	Cluster number, SF atom, atom energy, and mean energy	Cluster number, SF atom, atom energy, and mean energy	Cluster number, SF atom, atom energy, and mean energy
 1	<b>No. 2</b>	No. 7	No. 9	
	3. 3.64	8. 0.20	11. 0.19	
	2. 3.54	7. 3.61	10. 3.57	
	4. 2.95	9. 0.15	12. 0.14	
	6. 3.56			
	$\bar{E} = 3.42$	$\bar{E} = 1.32$	$\bar{E} = 1.3$	
 2	No. 7	<b>No. 15</b>	No. 18	No. 19
	2. 0.18	4. 3.62	8. 3.53	12. 3.58
	1. 3.52	1. 3.52	4. 3.62	11. 3.60
	3. 0.15	5. 6.64	9. 2.94	13. 0.00
	$\bar{E} = 1.28$	$\bar{E} = 4.33$	$\bar{E} = 3.42$	$\bar{E} = 2.7$
 97	<b>No. 2</b>	No. 207		
	10. 3.64	5. 0.04		
	1. 3.54	3. 3.59		
	11. 3.54	6. 0.03		
	12. 2.95	8. 0.03		
	$\bar{E} = 3.42$	$\bar{E} = 0.92$		

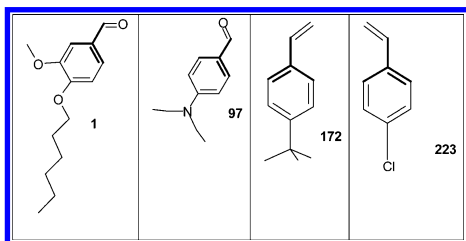


Figure 4. Cluster splitting example. The molecules part of one original cluster is divided into two new subsets.

MOA. In a previous work we selected 607 compounds; 10 compounds were excluded for practical reasons (e.g., they are salts). Demetra contains 282 compounds; no MOA determination is present in the database, and for very few compounds a MOA can be found in the literature. This second set has been selected because it is composed by highly varied and complex molecules; it represents a very complex problem. Before discussing our results, we present some examples of the four procedures described in the previous section, i.e. SF-FG mapping, cluster merging, cluster selection, and cluster splitting, in order to better illustrate our approach.

**3.1. Example of SF-FG Mapping.** In this example (Figure 1) all the SF atoms are contained in the FG, where the most important atom is 2 and the FG includes the nitro group and the ortho positions of the aromatic ring.

In this example (Figure 2) no FG contains all the SF atoms; in fact, the most important atoms are 2 and 4, and the FGs include some of the neighboring atoms but never 2 and 4 in the

same FG. Currently, we do not examine this compound further and all compounds in the same situation.

**3.2. Example of Cluster Merging.** As an example of merging we consider two clusters (7, 9) of Duluth; they contain, respectively, the compounds 45, 111, 224 and 100, 130, 163 (Figure 3).

Here, the two SFs are the aryl ether groups; however, the ACUs are different for methyl ethers and non-methyl ethers. The use of the FGs permits the determination of their high similarity without needing any a priori definition. A noteworthy side point concerns molecule 45; this is a very complex compound that is not easily inserted into the group of the others, but 45 remains in this cluster because it cannot find a better cluster and contains a common substructure that is wide enough (the aromatic ring with the three ether groups).

**3.3. Example of Cluster Selection.** In this example three molecules will be considered. It should be noted that it is not necessary to know the nature and composition of the used clusters, because the example uses a table of numbers that easily allows for the understanding of the selection made by the procedure.

Three molecules (1, 2, 97) are now examined to illustrate the selection of the best cluster.

Compound 1 belongs to three clusters: 2, 7, 9; compound 2 belongs to four clusters: 7, 15, 18, 19; compound 97 belongs to two clusters: 2, 207. In Table 1 are reported the compound structures, the energy weight of SF atoms, and the mean energy of each SF for each molecule in each cluster. As it can be seen, the best cluster is no. 2 for molecules 1 and 97 and no. 15 for

Table 2. Comparison of the Cluster Results with the MOA of the Duluth Data<sup>7</sup>

SF	number of molecule	Duluth MOA <sup>a</sup>	coverage <sup>b</sup>
2	28	17 electrophile, 9 narcotic, 2 undetermined	65%
15	28	11 narcotic, 8 electrophile, 5 uncoupler, 1 AChE inhibitor, 3 undetermined	44%
19	40	22 narcotic, 4 uncoupler, 3 AChE inhibitor, 2 electrophile, 1 CNS seizure, 8 undetermined	69%
36	8	4 narcotic, 4 undetermined	100%
42	8	7 narcotic, 1 electrophile	87%
45	19	16 narcotic, 2 electrophile, 1 undetermined	89%
56	5	4 neurodepressant, 1 undetermined	100%
76	24	19 narcotic, 2 CNS seizure, 2 electrophile, 1 undetermined	83%
87	4	2 narcotic, 1 electrophile, 1 undetermined	67%
90	7	7 narcotic	100%
92	25	18 narcotic, 1 neurodepressant, 1 CNS seizure, 1 electrophile, 4 undetermined	86%
130	20	19 narcotic, 1 electrophile	95%
152	5	2 narcotic, 2 AChE inhibitor, 1 undetermined	50%
156	7	4 narcotic, 1 electrophile, 2 undetermined	80%
173	4	1 narcotic, 3 undetermined	100%
206	21	13 narcotic, 1 electrophile, 7 undetermined	93%
226	3	2 narcotic, 1 electrophile	67%
237	15	2 narcotic, 13 undetermined	100%
252	17	11 narcotic, 1 AChE inhibitor, 5 undetermined	92%
253	8	8 undetermined	n.d.
300	8	7 narcotic, 1 undetermined	87%
401	8	6 narcotic, 1 electrophile, 1 undetermined	86%
490	5	4 electrophile, 1 undetermined	100%
495	8	5 narcotic, 2 respiratory blocker, 1 undetermined	71%
575	6	4 electrophile, 2 undetermined	100%
total	331	71 undetermined	

<sup>a</sup>Duluth MOA: baseline narcosis, polar narcosis, arylate and ester narcosis, electrophile or proelectrophile, neurodepressant, uncoupler of oxidative phosphorylation, central nervous system seizure mechanisms, Acetylcholinesterase inhibition, respiratory blocker or inhibition. <sup>b</sup>The coverage is determined considering the MOA containing the greatest number of compounds for each cluster. The value is determined taking the ratio between the number of the compounds in the selected MOA and the total number of compounds in the cluster, excluding the compounds with undetermined MOA.

molecule 2. In the table the best SF is highlighted for each compound.

**3.4. Example of Cluster Splitting.** The last example concerns the division of a single cluster into smaller sets using the FGs. The original group contains 34 molecules, but to make the example clear, we reduced the number to four compounds that will be split into two subsets. The compounds are sketched in Figure 4.

The SF in this case is made up by one substituted aromatic carbon, bonded to two unsubstituted aromatic carbons and to one carbon carrying a double bond. In contrast, the FGs are different because they extend to the electronically connected neighbors: an oxygen or a carbon atom, respectively. The consequence is that the four compounds are separated into two subsets (1, 97 and 172, 223). This example also shows that the use of the FGs, that are electronically defined and that can be

Table 3. Comparison of the Cluster Results with the MOA of the Moro et al. Data<sup>8</sup>

SF	number of molecules	MORO MOA <sup>a</sup>	coverage <sup>b</sup>	MORO/EPA <sup>c</sup>
2	28	21 narcotic, 3 electrophile, 1 CNS seizure, 4 undetermined	87%	31%
15	28	20 narcotic, 5 electrophile, 1 AChE inhibitor, 2 undetermined	77%	48%
19	40	29 narcotic, 4 electrophile, 2 neurodepressant/CNS seizure, 5 undetermined	83%	84%
36	8	5 narcotic, 1 electrophile, 1 CNS seizure, 1 undetermined	71%	75%
42	8	7 narcotic, 1 undetermined	100%	87%
45	19	16 narcotic, 3 electrophile	84%	78%
56	5	3 narcotic, 2 electrophile	60%	75%
76	24	19 narcotic, 3 electrophile, 1 AChE inhibitor, 1 undetermined	83%	65%
87	4	2 narcotic, 2 electrophile	50%	67%
90	7	4 narcotic, 1 electrophile, 1 CNS seizure, 1 undetermined	67%	57%
92	25	17 narcotic, 4 electrophile, 1 uncoupler, 1 AChE inhibitor, 1 neurodepressant, 1 undetermined	71%	57%
130	20	11 narcotic, 6 electrophile, 3 undetermined	65%	60%
152	5	3 narcotic, 2 undetermined	100%	25%
156	7	5 narcotic, 1 uncoupler, 1 undetermined	83%	80%
173	4	4 narcotic	100%	100%
206	21	17 narcotic, 2 electrophile, 2 undetermined	89%	86%
226	3	2 narcotic, 1 CNS seizure	67%	33%
237	15	10 narcotic, 2 electrophile, 3 undetermined	83%	100%
252	17	10 narcotic, 2 electrophile, 2 uncoupler, 1 AChE inhibitor, 3 undetermined	71%	50%
253	8	4 narcotic, 2 electrophile, 1 uncoupler, 1 undetermined	57%	-
300	8	7 narcotic, 1 electrophile	87%	86%
401	8	6 narcotic, 1 CNS seizure, 1 electrophile	75%	57%
490	5	2 narcotic, 2 AChE inhibitor, 1 undetermined	50%	0%
495	8	6 narcotic, 2 undetermined	100%	43%
575	6	4 narcotic, 1 electrophile, 1 CNS seizure	67%	25%
total	331	34 undetermined		

<sup>a</sup>MORO MOA: baseline narcosis, polar narcosis, arylate and ester narcosis, electrophile or proelectrophile, neurodepressant, uncoupler of oxidative phosphorylation, central nervous system seizure mechanisms, Acetylcholinesterase inhibition, respiratory blocker or inhibition. <sup>b</sup>The coverage is determined considering the MOA containing the greatest number of compounds for each cluster. The value is determined taking the ratio between the number of the compounds in the selected MOA and the total number of compounds in the cluster, excluding the compounds with undetermined MOA. <sup>c</sup>Congruence ratio between Duluth and Moro MOAs.

considered as reaction sites (the function of FGs in chemical reactivity is illustrated in ref 22), validates also the SFs from this point of view.

**3.5. Analysis of the MOA.** After the merging and divide operations the number of survived clusters is limited, and they only contain some of the original compounds; more exactly, in Duluth 331 compounds (divided into 25 sets) are classified as 607 compounds, while in Demetra 145 compounds (divided

Table 4. Comparison of the Cluster Results with the MOA of the Toxtree Data<sup>26</sup>

SF	number of molecules	Verhaar I MOA <sup>a</sup>	coverage <sup>b</sup>	Verhaar II MOA	coverage <sup>b</sup>
2	28	15 class 3, 9 class 1, 3 class 2, 1 class 5	55%	28 class 3	100%
15	28	8 class 2, 4 class 3, 2 class 1, 1 class 4, 13 class 5	53%	7 class 2, 4 class 3, 1 class 4, 15 class 5	54%
19	40	13 class 1, 6 class 4, 5 class 2, 3 class 3, 13 class 5	48%	11 class 1, 5 class 2, 2 class 3, 7 class 4, 15 class 5	44%
36	8	8 class 5	n.d.	8 class 5	n.d.
42	8	6 class 3, 2 class 5	100%	6 class 3, 2 class 5	100%
45	19	14 class 1, 1 class 3, 4 class 5	93%	14 class 1, 1 class 3, 4 class 5	93%
56	5	5 class 5	n.d.	5 class 5	n.d.
76	24	6 class 2, 1 class 3, 17 class 5	86%	6 class 2, 1 class 3, 17 class 5	86%
87	4	3 class 1, 1 class 3	75%	3 class 1, 1 class 3	75%
90	7	7 class 1	100%	7 class 1	100%
92	25	20 class 1, 1 class 3, 4 class 5	95%	20 class 1, 1 class 3, 4 class 5	95%
130	20	11 class 2, 9 class 5	100%	11 class 2, 9 class 5	100%
152	5	2 class 4, 3 class 5	100%	2 class 4, 3 class 5	100%
156	7	4 class 1, 3 class 5	100%	4 class 1, 3 class 5	100%
173	4	1 class 1, 3 class 5	100%	1 class 1, 3 class 5	100%
206	21	13 class 1, 8 class 5	100%	12 class 1, 1 class 3, 8 class 5	92%
226	3	3 class 1	100%	2 class 1, 1 class 3	67%
237	15	10 class 2, 1 class 3, 4 class 5	91%	11 class 2, 4 class 5	100%
252	17	1 class 1, 1 class 4, 15 class 5	50%	1 class 1, 1 class 4, 15 class 5	50%
253	8	5 class 1, 3 class 5	100%	5 class 1, 3 class 5	100%
300	8	8 class 5	n.d.	8 class 5	n.d.
401	8	2 class 1, 6 class 3	75%	8 class 3	100%
490	5	4 class 3, 1 class 5	100%	4 class 3, 1 class 5	100%
495	8	1 class 3, 7 class 5	100%	1 class 3, 7 class 5	100%
575	6	6 class 3	100%	6 class 3	100%
total	331				

<sup>a</sup>MOA Verhaar: 1 narcotic, 2 unreactive, 3 unspecifically reactive, 4 specifically reactive, 5 unclassified. Toxtree uses different calculation methods; in particular, two different methods are referred to as Verhaar (I) and Verhaar modified (II). <sup>b</sup>The coverage is determined considering the MOA containing the greatest number of compounds for each cluster. The value is determined taking the ratio between the number of the compounds in the selected MOA and the total number of compounds in the cluster, excluding the compounds with undetermined MOA.

into 23 sets) are classified as 282. It is worth remembering that Duluth contains 63 outliers and Demetra contains 36 outliers. The remaining missing compounds are in too small sets or have not been validated by SF-FG mapping. The comparison to the available experimental data is reported in Tables 2–5 for both of the databases. It is worth remembering that the experimental data for Duluth mainly come from EPA, while no experimental data are available for Demetra.

**3.5.1. Duluth Database.** The results show that the classification for Duluth is overall good, showing percentage agreement between 65 and 100 in all cases but one (SF 15). However, a deeper analysis of some selected cases can be more interesting. We have intentionally selected three cases: SF 2 class, SF 15 class, and SF 253 class.

The class contains 28 molecules; they are sketched in Figure 5.

The MOA classification by EPA gives 17 electrophiles, 9 narcotics, 2 undetermined; the best agreement is with electrophiles (65%).

The MOA classification by Toxtree (Verhaar I scheme) gives 15 compounds in class 3, 9 in class 1, 3 in class 2, 1 in class 5; the best agreement is with class 3 (55%).

The MOA classification by Toxtree (Verhaar II scheme) gives 28 compounds in class 3; the best agreement is with class 3 (100%).

The MOA classification by Moro gives 21 narcotics, 3 electrophiles, 1 CNS seizure, 4 undetermined; the best agreement is with narcotics (87%).

However, the agreement between Moro and EPA classification is low (31%).

This SF is characterized by the presence of an aromatic aldehyde that gives a reason for MOAs mainly pre-electrophiles/electrophiles with an intermediate confidence level. What about the misclassified compounds? Looking at the structures some special situations appear. For example, compounds **606**, **607**, **610**, and **615** are all very similar, but **606** is EPA classified as a narcotic. **606** and **607** differ for a CH<sub>2</sub> group, only; **607** and **1** differ for a longer alkyl chain, but **1** and **607** are nevertheless jointly classified as electrophiles. The FG is always the same (Figure 6).

**631**, **632**, **845**, and **858** are all salicylic aldehydes; they differ for the presence of Cl or Br atoms. However **631** and **632** are jointly classified, while **845** and **858** are not classified. Surprisingly, 4-isopropyl benzaldehyde (**802**) and 2-methyl benzaldehyde (**934**) are classified as narcotics, while **806** is electrophile; the confidence level for **806** is high, for **802** and **934** it is medium. Two subgroups (**97**, **217**, **918**; **232**, **234**, **909**) are somehow different: in the first the compounds have a tertiary amine substituent; in the second they carry a phenyl ether group. These substituents could support a different MOA, but in the classification this difference goes unnoticed.

Moro classification assigns the greatest part of the compounds to the narcosis class, limiting to two the number of electrophiles. The two Toxtree models show contrasting results: the first classifies as narcotics all the compounds that are EPA electrophiles and as reactive all the compounds that

Table 5. Comparison of the Cluster Results with the MOA of the Toxtree Data

SF	number of molecules	MOA Verhaar I <sup>a</sup>	coverage <sup>b</sup>	MOA Verhaar II	coverage <sup>b</sup>
1	6	2 class 2, 2 class 4, 2 class 5	50%	2 class 2, 2 class 4, 2 class 5	50%
9	7	7 class 5	n.d.	7 class 5	n.d.
17	11	2 class 4, 1 class 2, 8 class 5	67%	2 class 4, 1 class 2, 8 class 5	67%
31	6	1 class 3, 1 class 4, 4 class 5	50%	1 class 3, 1 class 4, 4 class 5	50%
36	4	4 class 1	100%	4 class 1	100%
50	12	1 class 4, 11 class 5	100%	1 class 4, 11 class 5	100%
55	5	3 class 3, 1 class 4, 1 class 5	75%	3 class 3, 1 class 4, 1 class 5	75%
93	8	5 class 4, 1 class 1, 1 class 3, 2 class 5	83%	4 class 4, 1 class 1, 2 class 3, 2 class 5	67%
99	10	3 class 4, 1 class 3, 6 class 5	75%	3 class 4, 1 class 3, 6 class 5	75%
104	4	1 class 1, 1 class 4, 2 class 5	50%	1 class 1, 1 class 4, 2 class 5	50%
110	7	3 class 1, 1 class 3, 3 class 5	75%	3 class 1, 1 class 3, 3 class 5	75%
122	7	2 class 3, 5 class 5	100%	2 class 3, 5 class 5	100%
138	6	6 class 4	100%	5 class 4, 1 class 3	83%
161	7	7 class 4	100%	7 class 4	100%
210	6	1 class 3, 5 class 5	100%	1 class 3, 5 class 5	100%
212	4	1 class 1, 1 class 2, 2 class 5	50%	1 class 1, 1 class 2, 2 class 5	50%
217	5	5 class 4	100%	5 class 4	100%
296	4	4 class 5	n.d.	4 class 5	n.d.
311	5	5 class 5	n.d.	5 class 5	n.d.
325	6	1 class 4, 5 class 5	100%	1 class 4, 5 class 5	100%
357	5	3 class 3, 2 class 5	100%	3 class 3, 2 class 5	100%
443	4	2 class 4, 2 class 5	100%	2 class 4, 2 class 5	100%
568	6	5 class 4, 1 class 3	83%	6 class 3	100%
total	145				

<sup>a</sup>MOA Verhaar: 1 narcotic, 2 unreactive, 3 unspecifically reactive, 4 specifically reactive, 5 unclassified. Toxtree uses different calculation methods; in particular, two different methods are referred to as Verhaar (I) and Verhaar modified (II). <sup>b</sup>The coverage is determined considering the MOA containing the greatest number of compounds for each cluster. The value is determined taking the ratio between the number of the compounds in the selected MOA and the total number of compounds in the cluster, excluding the compounds with undetermined MOA.

are EPA narcotics; the second classifies all the compounds as reactive, in agreement with our model.

Curiously, compounds **606** and **607** show two diverse MOAs for the first Toxtree model, but both are narcotics for the Moro model. This example accurately illustrates the limits of the experimental data and the difficulties that models find because of these limits.

The class contains 28 molecules; they are sketched in Figure 7.

The MOA classification by EPA gives 11 narcotics, 8 electrophiles, 5 uncouplers, 1 AChE inhibitor, 3 undetermined; the best agreement is with narcotics (44%).

The MOA classification by Toxtree (Verhaar I scheme) gives 2 compounds in class 1, 8 in class 2, 4 in class 3, 1 in class 4, 13 in class 5; the best agreement is with class 2 (53%).

The MOA classification by Toxtree (Verhaar II scheme) gives 7 compounds in class 2, 4 in class 3, 1 in class 4, 15 in class 5; the best agreement is with class 2 (54%).

The MOA classification by Moro gives 20 narcotics, 5 electrophiles, 1 AChE inhibitor, 2 undetermined; the best agreement is with narcotics (77%).

However, the agreement between Moro and EPA classification is again low (48%).

This SF is mainly formed by nitrobenzenes; the nitro group is highly characterizing thus hiding the structure differences between the molecules; thus, it collects compounds that have different MOAs, among which 39% are narcotics. Simple nitrobenzenes with or without an OH or a NH<sub>2</sub> group are all narcotics or polar narcotics; dinitrobenzenes and aldehydes are electrophiles; dinitrobenzenes with an OH group are uncouplers; **178** is an AChE inhibitor. This example clearly

shows that the choice of the best SF to group compounds can give erroneous results. The nitro group is structurally important; however, it seems that the presence of other groups is particularly significant. This problem should be considered in future developments.

The agreement is low also for the other models, as can be seen in the tables. In the Moro model most of the compounds are narcotics. In addition, the AChE inhibitor is compound **619** that is very similar to both **616** and **621**. The Toxtree model cannot assign a classification to most of the compounds.

The class contains 8 molecules; they are sketched in Figure 8.

The MOA classification by EPA gives 8 compounds undetermined.

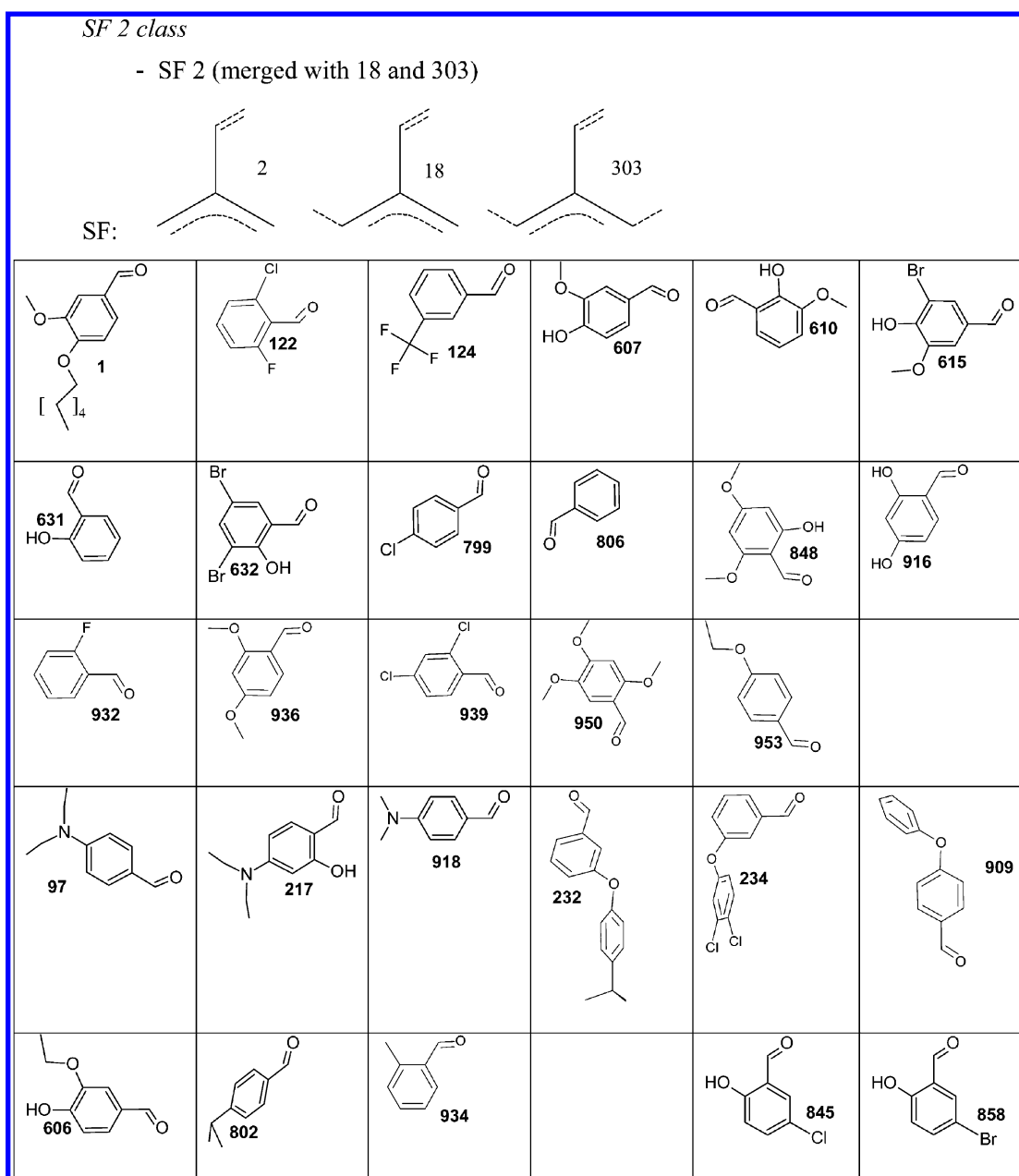
The MOA classification by Toxtree (Verhaar I scheme) gives 5 compounds in class 1, 3 in class 5; the best agreement is with narcotics (100%).

The MOA classification by Toxtree (Verhaar II scheme) gives 5 compounds in class 1, 3 in class 5; the best agreement is with narcotics (100%).

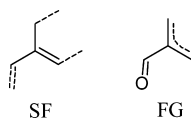
The MOA classification by Moro gives 5 narcotics, 2 electrophiles, 1 undetermined; the best agreement is with narcotics (71%).

This SF is characterized by the presence of the carboxyl group. It contains two clear subgroups: the first including compounds **88**, **833**, **876**, and **958** contains simple alkyl derivatives; the second including compounds **129**, **134**, **171**, and **201** contains complex structures. Quite surprisingly, the EPA classification is absent for all the compounds. The Toxtree model assigns in both versions some compounds to the narcosis class, being unable to classify the others (**129**, **134**, and **201**). The Moro model, vice versa, gives more articulated





**Figure 5.** Structural fragment 2 class. The cluster includes molecules of the original 2, 18, and 303 classes. The molecule MOAs are different depending on the considered data.



**Figure 6.** Structural fragment and FG of the SF 2 class.

answers; apart of 5 narcotics, the model also determines two electrophiles (**88** and **958**): this result is unexpected because these two compounds are simple alkyl acids and there is no reason to believe that they can act as electrophiles. The structures of the compounds are really different; however, the carboxyl group can be considered the dominating functional group, and our result only reflects this fact.

**3.5.2. Demetra.** The results show that the classification for Demetra is apparently good, showing percentage agreement between 50 and 100 in all the cases. However, the comparison

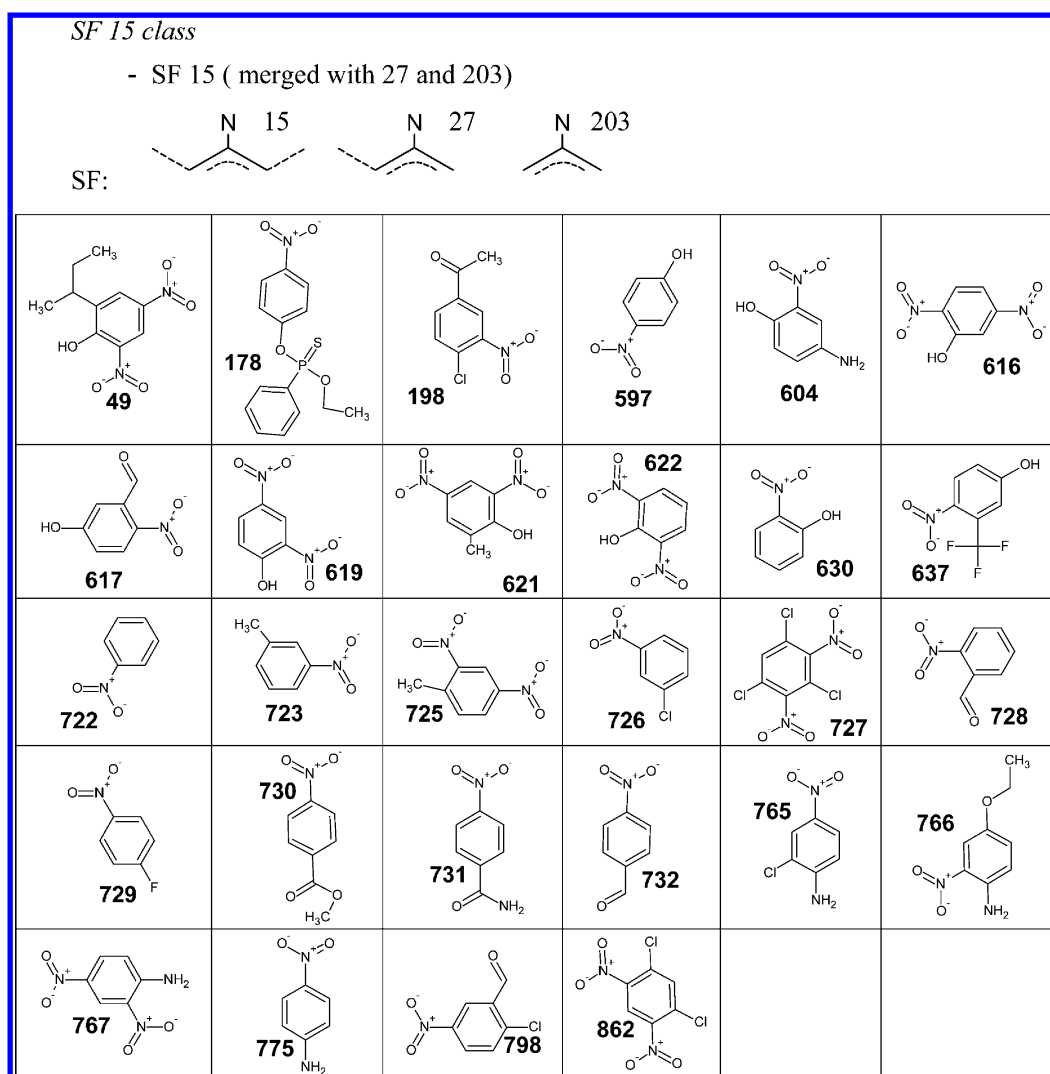
is done using only the Toxtree model, thus its reliability is limited. In fact, looking at the number of classified compounds we can note that most of them are not analyzed by Toxtree. Nevertheless, we are going to give a deeper analysis of some selected cases that we consider worth discussing. We have selected four cases: SF 93 class, SF 138 class, SF 161 class, and SF 568 class.

The class contains 9 molecules; they are sketched in Figure 9.

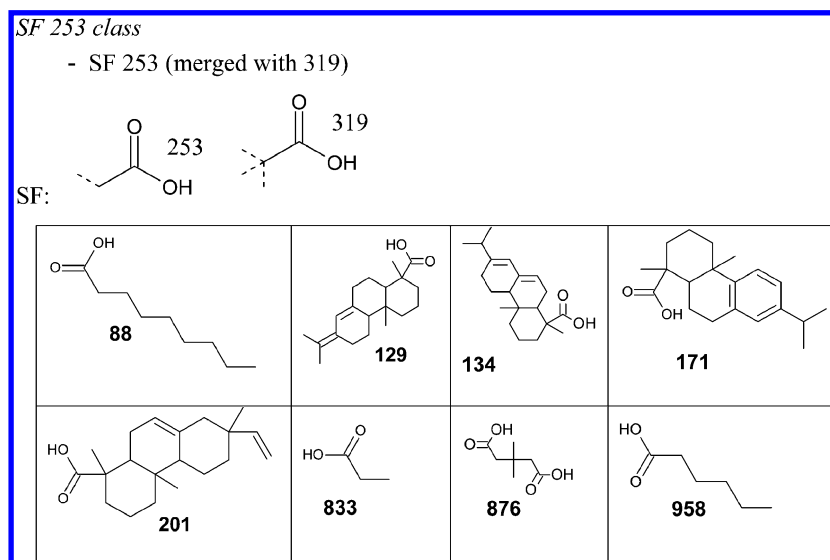
The MOA classification by Toxtree (Verhaar I scheme) gives 1 compounds in class 1, 1 in class 3, 4 in class 4, 3 in class 5; the best agreement is with reactivities (67%).

The MOA classification by Toxtree (Verhaar II scheme) gives 1 compounds in class 1, 2 in class 3, 3 in class 4, 3 in class 5; the best agreement is with reactivities (50%).

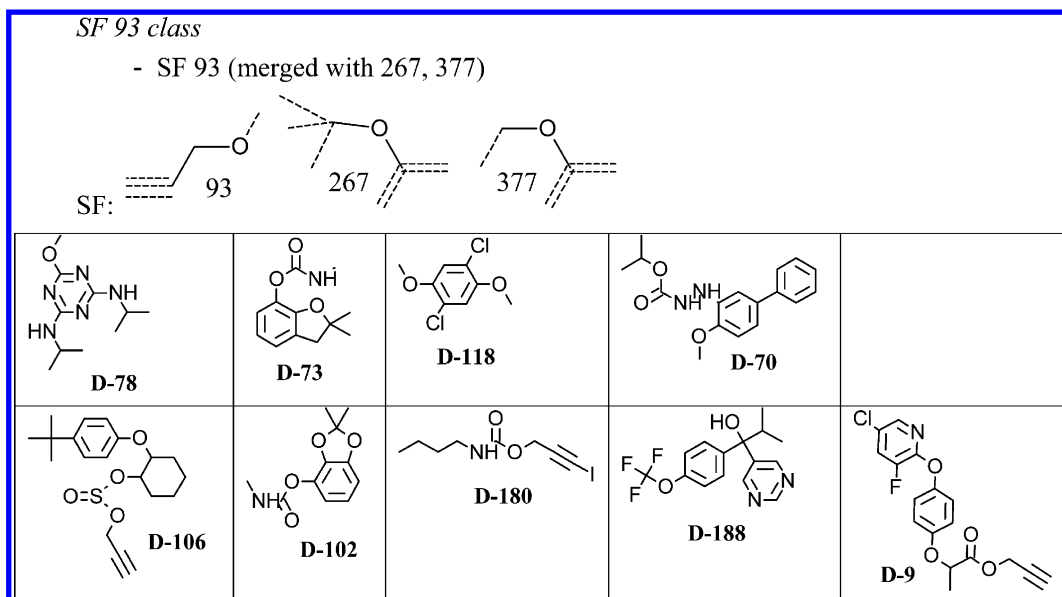
In this SF there are three subgroups containing compounds **D-78**, **D-73** + **D-118** + **D-70**, and **D-106** + **D-9** + **D-102** + **D-**



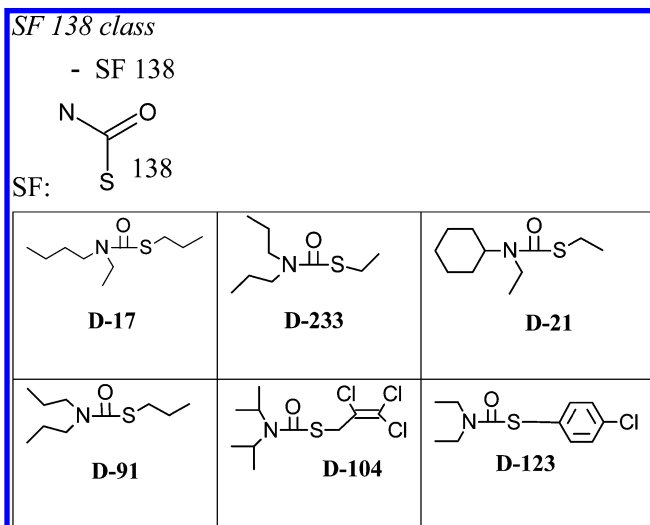
**Figure 7.** Structural fragment 15 class. The cluster includes molecules of the original 15, 27, and 203 classes. The molecule MOAs are different depending on the considered data.



**Figure 8.** Structural fragment 253 class. The cluster includes molecules of the original 253 and 319 classes. The molecule MOAs are different depending on the considered data.



**Figure 9.** Structural fragment 93 class. The cluster includes molecules of the original 93, 267, and 377 classes. The molecule MOAs are different as the corresponding structures.



**Figure 10.** Structural fragment 138 class. The cluster includes molecules of the original 138 class only. The molecule MOA is common to all compounds.

180 + D-188, respectively. The first point concerns the participation of compounds containing a triple bond to a group of arenes: but the energies of the two bond types are computationally very similar. The classes (3 and 4) of reactive compounds are the most populated for both Toxtree models; compound D-118 is the only narcotic, in agreement with its simple structure. The two molecules carrying a triple bond (D-106 and D-180) are not classified, while compound D-9 (also carrying a triple bond) is classified reactive. The presence of the subgroups does not seem significant. Everything considered the group is more homogeneous than expected.

The class contains 6 molecules; they are sketched in Figure 10.

The MOA classification by Toxtree (Verhaar I scheme) gives 6 compounds in class 4; the best agreement is with reactives (100%).

The MOA classification by Toxtree (Verhaar II scheme) gives 1 compound in class 3, 5 in class 4; the best agreement is with reactives (83%).

This SF contains compounds belonging to the same chemical group: carbamates. It is very homogeneous; in fact, both Toxtree models give a clear classification. In the second model compound D-104 is differentiated from the others; nevertheless it remains in the reactive general area. The FG analysis is highly consistent with the classification.

The class contains 7 molecules; they are sketched in Figure 11.

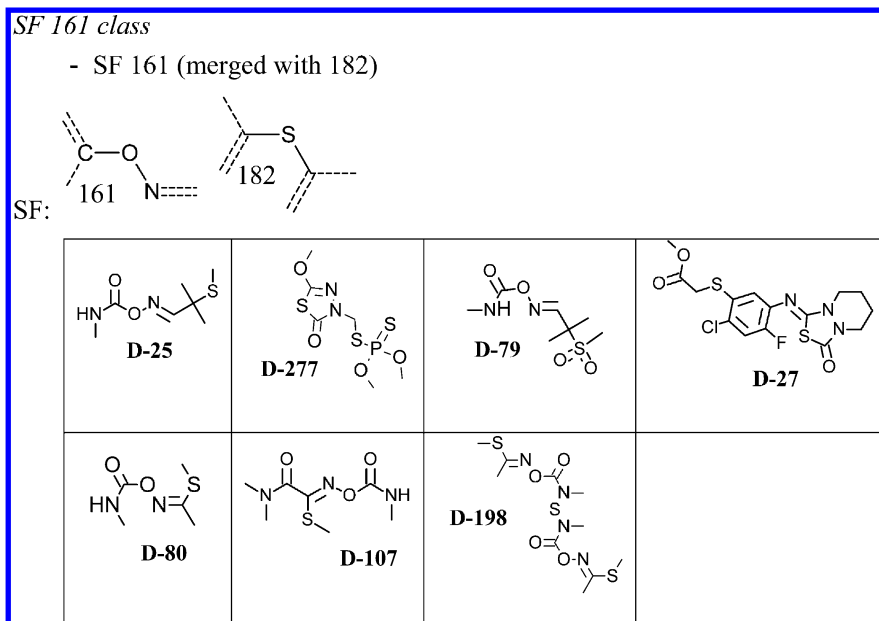
The MOA classification by Toxtree (Verhaar I scheme) gives 7 compounds in class 4; the best agreement is with reactives (100%).

The MOA classification by Toxtree (Verhaar II scheme) gives 7 compounds in class 4; the best agreement is with reactives (100%).

This group contains highly functionalized compounds that are more or less similar. Three compounds (D-25, D-80, and D-107) are present also in the Duluth database where they are classified as AChE inhibitors. Even considering that the Duluth data concern a different fish, it makes sense to suggest that all the compounds should have the same MOA. However, there is one problem: the analysis of the FG shows that the SF of the group is not correctly represented by the corresponding FG (Table 6); as a consequence the group is out of the procedure. In fact, the group formation is based on the SFs and the FGs are used only to merge and divide the original groups in order to have groups validated by the energetic control; thus, if the SF atoms are not completely contained in one FG the group membership of the molecule is not validated. If the validation is missing for all the molecules in the group, then this is discarded.

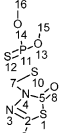
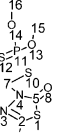
The class contains 6 molecules; they are sketched in Figure 12.

The MOA classification by Toxtree (Verhaar I scheme) gives 1 compound in class 3, 5 compounds in class 4; the best agreement is with reactives (83%).



**Figure 11.** Structural fragment 161 class. The cluster includes molecules of the original 161 and 182 classes. The molecule MOA is common to all compounds.

**Table 6.** Comparison of the SF and FG Atoms for Two Molecules of the Set<sup>a</sup>

	SF		FG 1		FG 2		SF		FG 1		FG 2
	O	3	O	8	N	4		S	1	C	5
	C	2	C	2	C	5		C	2	O	8
	N	4	N	1	O	3		C	5	N	4
			O	3	C	6			S	1	O
			C	10	S	7				S	1
										C	9
D-25							D-277				

<sup>a</sup>SF and FG atoms of compounds **D-25** and **D-277**. The SF atoms are not concurrently present in any FGs.

The MOA classification by Toxtree (Verhaar II scheme) gives 6 compounds in class 4; the best agreement is with reactivities (100%).

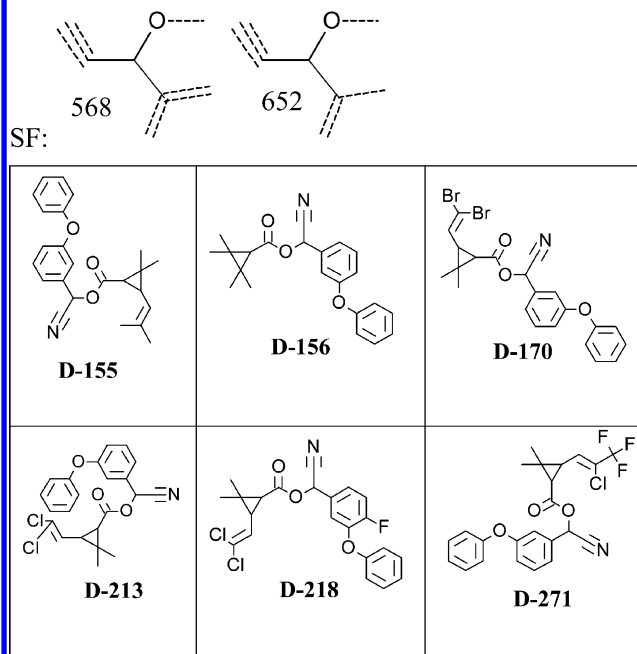
This SF is highly homogeneous containing only pyrethroids. Thus, one common MOA is expected. The two Toxtree models, in fact, give similar classification; only model I separates compound **D-156**, assigning it to a different reactive class. Here, the characteristic structure helps in collecting compounds with a similar MOA; it should be pointed out that the result could have been different in a different database.

#### 4. DISCUSSION

Often, when studying the environmental toxicity of chemicals the causes are not sufficiently detailed, and it is not uncommon that the experimental data are limited to black and white alternatives: the animal, or plant, lives or dies. The current availability of experimental data is not sufficient to reach the molecular level and, then, to determine the MOA. This severely limits all modeling attempts to assist the analysis.

A possible solution is to perform the study starting from the fundamentals of chemistry. The hypothesis is that MOAs are common reactions between two reactants: the first the well-

**SF 568 class**  
 - SF 568 (merging with 652)



**Figure 12.** Structural fragment 568 class. The cluster includes molecules of the original 568 and 652 classes. The molecule MOA is common to all compounds.

known chemical compound; the second the ill-defined biological target. Even if the second partner is not well-known, and the reaction depends on both parts, it is sufficiently consistent that the reactivity of the first partner does not change very much. Thus, we aim at developing a model based on chemical reactivity that can be used to group compounds with similar reactivity and, perhaps, similar MOA.



The possibility to develop predictive models to assist MOA recognition is consistent with the recommendations made by the regulators. Interpreting the MOA as a reaction preference, we can state that our model groups together all compounds that have a similar reactivity. If we add to this the requirement that the compounds in a group are also structurally similar, we can expect a result showing more groups than MOAs, but the groups contain similar molecules sharing a common MOA. The groups can be now analyzed against the known MOAs.

Here, an important problem emerges: the poor availability of experimentally determined MOAs for environmental toxicity. One data set that provides this information was compiled by EPA and concerns the toxicity of chemicals on fathead minnow (the well-known Duluth database). In principle, the cross check between the database and the clusters can validate the approach. However, the analysis gives highly varied suggestions. Some groups show good agreement with the EPA data; other groups are clearly not homogeneous enough to share a common MOAs. Our results also indicate that the experimental data are not consistent enough. This is clearly shown by the analysis of SF 2 group discussed in the Results section. Consequently, the model can be also used to analyze the experimental data. The comparison with other available models shows that all models currently miss a good standard result. Tables 2, 3, and 4 report the comparison with both experimental (Table 2) and model (Tables 3 and 4) determination of MOAs. It is possible to discuss the agreement of our results with these available data; nevertheless, it should be noted that our approach only classifies compounds without predicting any MOA; the MOA assignment is done by assuming that the group MOA is the MOA experimentally determined for the greatest number of compounds in the group. Experimentally defined MOAs (from Duluth database) are in 80% overall agreement with our results. This figure is comparable to that obtained by other models. Moro et al. model predictions are in 75% overall agreement with our results. The agreement of this model predictions with Duluth data (60%) makes our result satisfactory. ToxTree model predictions are in 86–89% overall agreement with our results. As already mentioned, no existing model can be considered reliable enough to classify toxic MOAs for chemicals. This is certainly due to the low quality and quantity of the experimental data and to the complexity of the problem. The approach presented here is a first step toward the development of a model independent of experimental data, very different from usual training – test sets procedures.

Attention should be drawn to the use of the obtained results: our strategy is limited to the formation of subgroups of compounds that share structural similarity from both atom-bond and electronic energy viewpoints; no MOA is assigned. Nonetheless, MOA assignment may eventually be performed by comparing the subgroups with experimentally determined MOA groups. It should be possible to use the obtained groups to assign theoretical MOA by adding other descriptors that can describe the compound reactivity. To this end, any kind of descriptors, including quantum chemical ones, could be exploited.

## 5. CONCLUSION

In this paper we have described an approach to classify compounds based on both a hierarchy of molecular fragments and their electronic description. The method does not use a priori knowledge and can be considered data-driven. The

discussion shows that the use of this unbiased analysis can give suggestions even for experimental data validation. It is nevertheless clear that the approach needs some improvements; in addition, due to the relatively scarce knowledge of the complex mechanisms involved in toxicity, a better and more extensive availability of data concerning MOAs is required.

## ■ APPENDIX

### Calculation of Electronic Energy

Relation between chemical potential and electronic energy

$$\mu = (\partial E / \partial N)_Z$$

where  $E$  is the electronic energy,  $N$  is the number of atom electrons, and  $Z$  is the core potential.

Definition of the chemical potential by Gordy approximation

$$\chi = k_1 \times Z_{\text{eff}} / r + k_2 = -\mu$$

where  $\chi$  is the atom electronegativity,  $Z_{\text{eff}}$  is the effective core potential as defined by Slater,  $r$  is the atom radius, and  $k_1$  and  $k_2$  are constants depending on the atom.

Dependence of atom covalent radius on the effective core potential of the nucleus

$$r(i) = r \times Z'_{\text{eff}} / Z'_{\text{eff}}(i)$$

where  $Z'_{\text{eff}}$  is the effective core potential of the nucleus shielded by all electrons of the isolated atom,  $r$  is the atom radius of the isolated atom, and  $Z'_{\text{eff}}(i)$  is the effective core potential of the nucleus shielded by all electrons at iteration  $i$ .

Calculation of the electronic energy by integration of the chemical potential with respect to the electron variation

$$E = \int \mu \partial N_z$$

$$E = k_3(A + B + C) - k_2 N_3$$

where  $A$ ,  $B$ ,  $C$ , and  $k_3$  are defined below

$$k_3 = -k_1 / (Z'_{\text{eff}} \times r)$$

where  $Z'_{\text{eff}}$  is the effective core potential of the nucleus shielded by all electrons of the isolated atom, and  $r$  is the atomic covalent radius of the isolated atom

$$A = (N^2 + aN - 2NN_1 - 2bNN_2 + N_1^2 + 2bN_1N_2 - aN_1 + b^2N_2^2 - abN_2)N_3$$

$$B = 0.5(-2aN + 2aN_1 + 2abN_2 - a^2)N_3^2$$

$$C = (a^{2/3})N_3^3$$

where  $N_1$ ,  $N_2$ , and  $N_3$  are the number of electrons in the first, second, and third shell, respectively.  $N$  is the total number of electrons.  $a$ ,  $b$ , and  $c$  are Slater's constants.

The calculation is iterative varying the number of third shell electrons and covalent radius until the difference of the energy of iteration  $n$  and iteration  $n-1$  is below a threshold.

When an atom is cut out of the molecule, no modification of bonds is made; the effect of the elimination of one atom only alters the chemical potential of the atoms that were connected to it.

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: guido.sello@unimi.it.

## Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The helpful contribution of G. S. Sinibaldi is gratefully acknowledged.

## ■ REFERENCES

- (1) European Commission, Environmental Directorate General (2007), (EC 1907/2006). <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32006R1907:EN:NOT> (accessed April 2012).
- (2) Guyton, K. Z.; Barone, S., Jr.; Brown, R. C.; Euling, S. U.; Jinot, J.; Makris, S. Mode of Action Frameworks: A Critical Analysis. *J. Toxicol. Environ. Health, Part B* **2008**, *11*, 16–31.
- (3) Hill, A. B. The Environment and Disease: Association or Causation? *Proc. R. Soc. Med.* **1965**, *58*, 295–300.
- (4) Netzeva, T. I.; Pavan, M.; Worth, A. P. Review of (Quantitative) Structure – Activity Relationships for Acute Aquatic Toxicity. *QSAR Comb. Sci.* **2008**, *27*, 77–90.
- (5) Gramatica, P.; Vighi, M.; Consolaro, F.; Todeschini, R.; Finizio, A.; Faust, M. QSAR Approach for the Selection of Congeneric Compounds with a Similar Toxicological Mode of Action. *Chemosphere* **2001**, *42*, 873–883.
- (6) (a) Raevsky, O. A.; Grigoreva, V. Y.; Weberb, E. E.; Dearden, J. C. Classification and Quantification of the Toxicity of Chemicals to Guppy, Fathead Minnow, and Rainbow Trout. Part 1. NonPolar Narcosis Mode of Action. *QSAR Comb. Sci.* **2008**, *27*, 1274–1281. (b) Raevsky, O. A.; Grigoreva, V. Y.; Weberb, E. E.; Dearden, J. C. Classification and Quantification of the Toxicity of Chemicals to Guppy, Fathead Minnow, and Rainbow Trout. Part 2. Polar Narcosis Mode of Action. *QSAR Comb. Sci.* **2009**, *28*, 163–174. (c) Raevsky, O. A.; Grigoreva, V. Y.; Tikhonova, O. V. Development of Structure-Toxicity Relationship Models of Chemicals with Respect to Guppy. *Pharm. Chem. J.* **2009**, *43*, 3125–3129.
- (7) [http://www.epa.gov/nct/dsstox/sdf\\_epafhm.html](http://www.epa.gov/nct/dsstox/sdf_epafhm.html) (accessed April 2012).
- (8) Michielan, L.; Pireddu, L.; Floris, M.; Moro, S. Support Vector Machine (SVM) as Alternative Tool To Assign Acute Aquatic Toxicity Warning Labels to Chemicals. *Mol. Inf.* **2010**, *29*, 51–64.
- (9) Lodhi, H.; Muggleton, S.; Sternberg, M. J. E. Multi-class Mode of Action Classification of Toxic Compounds Using Logic Based Kernel Methods. *Mol. Inf.* **2010**, *29*, 655–664.
- (10) Casalegno, M.; Sello, G.; Benfenati, E. Definition and Detection of Outliers in Chemical Space. *J. Chem. Inf. Model.* **2008**, *48*, 1592–1601.
- (11) Casalegno, M.; Sello, G.; Benfenati, E. Top-Priority Fragment QSAR Approach in Predicting Pesticide Aquatic Toxicity. *Chem. Res. Toxicol.* **2006**, *19*, 1533–1539.
- (12) Casalegno, M.; Benfenati, E.; Sello, G. Identification of Toxicifying and Detoxifying Moieties for Mutagenicity Prediction by Priority Assessment. *J. Chem. Inf. Model.* **2011**, *51*, 1564–1574.
- (13) Lozano, S.; Halm-Lemeille, M.-P.; Lepailleur, A.; Rault, S.; Bureau, R. Consensus QSAR Related to Global or MOA Models: Application to Acute Toxicity for Fish. *Mol. Inf.* **2010**, *29*, 803–813.
- (14) Escher, B. J.; Hermens, J. L. M. Modes of Action in Ecotoxicology: Their Role in Body Burdens, Species Sensitivity, QSARs, and Mixture Effects. *Environ. Sci. Technol.* **2002**, *36*, 4201–4217.
- (15) (a) Nendza, M.; Muller, M. Discriminating Toxicant Classes by Mode of Action 1. Physico-Chemical Descriptors. *Quant. Struct.-Act. Relat.* **2000**, *19*, 581–598. (b) Nendza, M.; Wenzel, A. Discriminating Toxicant Classes by Mode of Action 2. (Eco)toxicity Profiles. *Environ. Sci. Pollut. Res.* **2006**, *13*, 192–203.
- (16) Colombo, A.; Benfenati, E.; Karelson, M.; Maran, U. The Proposal of Architecture for Chemical Splitting To Optimize QSAR Models for Aquatic Toxicity. *Chemosphere* **2008**, *72*, 772–780.
- (17) Yuan, H.; Wang, Y.-Y.; Cheng, Y.-Y. Mode of Action-Based Local QSAR Modeling for the Prediction of Acute Toxicity in the Fathead Minnow. *J. Mol. Graphics Model.* **2007**, *26*, 327–335.
- (18) Ivanciuc, O. Support Vector Machines Prediction of the Mechanism of Toxic Action from Hydrophobicity and Experimental Toxicity Against *Pimephales promelas* and *Tetrahymena Pyriformis*. *Internet Electron. J. Mol. Des.* **2004**, *3*, 802–821.
- (19) Öberg, T. A QSAR for Baseline Toxicity: Validation, Domain of Application, and Prediction. *Chem. Res. Toxicol.* **2004**, *17*, 1630–1637.
- (20) Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A Comparative Study of Molecular Similarity, Statistical, and Neural Methods for Predicting Toxic Modes of Action. *Environ. Toxicol. Chem.* **1998**, *17*, 1056–1064.
- (21) Merlot, C. Computational Toxicology—a Tool for Early Safety Evaluation. *Drug Discovery Today* **2010**, *15*, 16–22.
- (22) Sello, G. A New Definition of Functional Groups and a General Procedure for Their Identification in Organic Structures. *J. Am. Chem. Soc.* **1992**, *114*, 3306–3311.
- (23) Casalegno, M.; Sello, G.; Benfenati, E. Top-Priority Fragment QSAR Approach in Predicting Pesticide Aquatic Toxicity. *Chem. Res. Toxicol.* **2006**, *19*, 1533–1539.
- (24) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity Using 2D Fragment Bit-Strings. *Comb. Chem. High Throughput Screening* **2002**, *5*, 155–166.
- (25) Bader, R. F. W. Atoms in Molecules. *Acc. Chem. Res.* **1985**, *18*, 9–15.
- (26) Toxtree (Estimation of Toxic Hazard - A Decision Tree Approach), Ideconsult Ltd., Version 2.5.0. <http://toxtree.sourceforge.net> (accessed April 2012).