Definition and Detection of Outliers in Chemical Space[#]

Mosè Casalegno,[*,†] Guido Sello,[‡] and Emilio Benfenati[†]

IRFMN, Mario Negri Institute for Pharmacological Research, Via La Masa, 19, 20156 Milano, Italy, and
Dipartimento di Chimica Organica e Industriale, Università degli Studi di Milano,
Via Venezian 21, 20133 Milano, Italy

Received November 7, 2007

Detection of outliers is a complex and challenging area of research in chemical theory. Among current notions, that of outliers in the chemical space—descriptors—is meaningful with multiple applications in the field of drug discovery and predictive modeling. Presented here is a new framework for outlier detection, relying on a discrete, fragment-based representation of the molecular structures. From this starting point, a recursive method is developed that quantifies the contribution of fragments to compound description and identifies outliers in chemical structure databases according to a novel definition. In contrast to existing detection routes, this approach avoids the use of thresholds usually required to quantify outlying behavior. Three chemical databases are investigated to demonstrate its generality and flexibility. The result reveals a new species of outliers, compounds with no specific structural features, rather than unique ones.

1. INTRODUCTION

Over the years, outlier detection has been recognized as an important and challenging task in chemical theory. The term outlier generally refers to a compound with one or more properties or attributes that substantially differ from others. This definition is general enough to encompass most of those currently in use and qualitatively introduces the topic of outlier detection. However, the literature still offers no universally accepted definition of outlier. Different notions have been proposed and adopted within different research fields.[1] One meaningful notion of outlier is associated with analysis of the chemical space. The chemical space, interpreted as descriptors space,[2] is the multidimensional space generated by a set of attributes—descriptors—used to characterize a chemical structures database. Mining outliers in the chemical space means identifying the compounds that are structurally different from the rest of the molecules in the database, with regards to the set of descriptors used. The search for such points in the chemical space has many potential applications in drug discovery[3–6] and predictive modeling.[7–13]

The notion of outlier in the descriptors space is now receiving increasing attention in the context of model applicability. Former approaches to outlier detection were mainly concerned with the analysis of the response space. The presence of outliers was based on a posteriori examination of the prediction error. Outlier detection was typically done after prediction, by comparing predicted and observed responses for each compound queried. Outliers were compounds showing exceptional prediction errors. They were analyzed with the aim of clarifying the mechanisms of toxic action of some chemicals.[14–16] In other cases, prediction outliers were removed in order to improve a model's reliability[16] or to judge the relative importance of descriptors within a specific model.[17]

Recently, however, the need to supply experimental data for a huge number of industrial chemicals,[18] and to speed up decision-making in risk assessment, has renewed interest in methods to identify outliers.[7] This research direction reflects the fact that experimental data are not normally available for new query compounds, and thus the a posteriori presence of outliers cannot be assessed. Instead, one can characterize the description of the chemical space spanned by the molecules used to build the model, the training set, on the assumption that a query compound well described by the training set molecules should be predicted more reliably than a dissimilar one.[8] The validity of this assumption has been tested in many studies, most of them aimed at estimating a model's applicability domain.[8–13] It has been shown[9–11] that the prediction error for query compounds outside the training set domain is, on the average, larger than for compounds within the domain. This does not necessarily imply that the prediction for a suspected outlier would be incorrect.[10] It does mean, however, that the estimate is less reliable and must be carefully evaluated before being accepted.

Many approaches have been proposed to measure compounds diversity in the context of diversity analysis[2,6,19–26] and applicability domain estimation.[10,12,27–31] Although most of them could, in principle, be used to detect outliers, only a few do explicitly deal with this issue.[2,6,12] What is common to all these methods, regardless of the specific application area, is the use of predefined thresholds to distinguish between outliers and nonoutliers. Thresholds are needed because we will not be able to detect outliers unless we first quantify outlyingness, by setting or calculating appropriate cutoffs. In common practice, outlier detection is usually repeated for different cutoff values, until a suitable setting is found.[2,32] This approach does not limit applicability of the method; however, it does require the user to put

[*] Corresponding author e-mail: casalegno@marionegri.it.
[†] Mario Negri Institute for Pharmacological Research.
[‡] Università degli Studi di Milano.
[#] Dedicated to E. J. Corey on the occasion of his 80th birthday.

10.1021/ci7004065 CCC: $40.75
Published on Web 07/25/2008

considerable trial-and-error effort into the search, which is impractical for large-scale application. In addition, it cannot provide general recommendations that could be used in passing from one database to another.

Here, we describe a fragment-based recursive clustering scheme for outlier detection that avoids the need for threshold parameters. In traditional clustering, the objects are collected into clusters that are initially void and form applying a search method; in contrast, our method starts from a predefined set of clusters, whose contributions to compound description are iteratively calculated until their final values are obtained.

At the base of our method is a framework where molecular structures are represented by means of fuzzylike memberships,[33] and a similarity function, called affinity, quantifies the tendency of a compound to belong to a cluster. Along with the adoption of this framework, a novel definition of outlier emerges that provides a numerical criterion for outlier detection. Once the convergence is reached, this criterion is applied to identify poorly described compounds, without the need for cutoffs.

In the present work, our aim is to describe an alternative strategy for outlier detection that applies to arbitrary discrete descriptor spaces. As an application, we consider three chemical structure databases differing in size and structural diversity.

2. METHODS

2.1. Overview. Here we provide a qualitative view of our approach. Our route to outlier detection begins with characterization of the molecular structures on the basis of structural features, fragments. Like other molecular descriptors, fragments generate a chemical space in which to search for outliers. Mining this space directly by considering the diversity in terms of distance between compounds, for example, would necessarily require us to explicitly set thresholds to isolate outliers, which is not our aim. We can, however, exploit the discrete nature of such descriptors to develop a cluster-based approach where outliers can be detected in an indirect way.

The basic idea is to map the fragment-based representation onto a cluster-based one. Clusters are groups of compounds sharing a common fragment. As a consequence, each compound belongs to as many clusters as the number of its constituent fragments. This introduces a descriptive framework where compounds are described by means of membership in clusters (also referred to as sets), and a function called affinity quantifies the tendency of a compound to belong to a cluster.

The method we describe is iterative, as it attempts to solve a problem by finding successive approximations to the solution, starting out from a guess. In this case, the problem we want to solve is quantifying the contribution of each cluster to compounds description.

The described approach does not locate outlying compounds. We, therefore, introduce a simple expedient that transforms it into a tool for outlier detection. We assign to each molecule an additional artificial set, called outlier set. The outlier set quantifies the extent to which a compound is not described by the other sets. Consequently, we define an outlier as a compound having unitary membership in the outlier set.

This definition provides a simple criterion to numerically characterize outliers. Once the iterative procedure has reached convergence, the criterion is applied to all compounds. A unitary membership in the outlier set is a sufficient condition for a compound to become an outlier.

2.2. Mathematical Framework. In this section, we outline the mathematical framework of the method and introduce the entities needed to implement the outlier detection algorithm. We consider a chemical structure database containing N compounds. The route begins with characterization of all the compounds using 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 × M elements, is filled as follows:

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

The occupancy matrix provides the first input to the algorithm described in this section. The second input is given by a pairwise similarity matrix S, consisting of N × N elements, whose generic element S(i,ii) quantifies the degree of structural resemblance between the i-th and ii-th molecules, defined as follows

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

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]. As we shall see below, the pairwise similarity matrix serves to drive the recursive evaluation of the clustering process.

At this point, we introduce the conceptual framework that is the basis of our approach. First, we count the total number of molecules where the j-th fragment occurs, n(j), defined by the following equation:

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

Then, we associate with each fragment, a set C(j), defined as the collection of n(j) molecules where the j-th fragment occurs, i.e. all compounds for which O(i,j) = 1. The meaning of C(j) is easy to understand. C(j) identifies the cluster of compounds that share one common fragment, the j-th. Like in the previous section, we use the terms cluster and set interchangeably. The fragments for which n(j) = 1 correspond to the singly occurring fragments (SOFs) encountered above. The sets associated with singly occurring fragments contain only one compound and will not be further considered. Our aim is to describe all molecular structures by means of the clusters C(j) so we need to mathematically quantify the relationship between molecules and clusters. 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:

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

A typical example of a clustering scheme resulting from this strategy is given in Figure 1. The compound of interest (4-

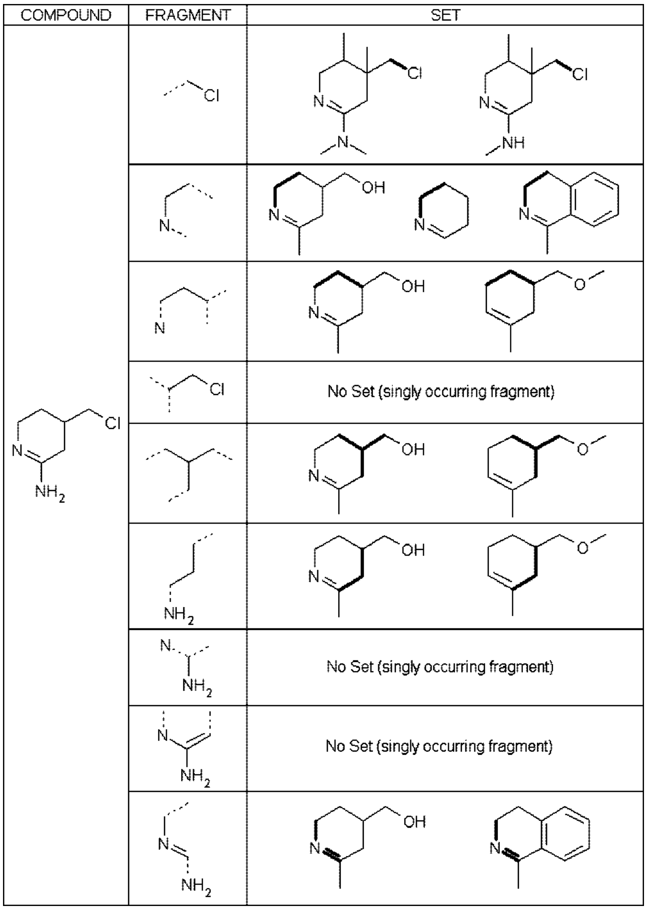

COMPOUND
FRAGMENT
SET
No Set (singly occurring fragment)
No Set (singly occurring fragment)
No Set (singly occurring fragment)

Figure 1. Example of a clustering scheme for 4-chloromethyl-3,4,5,6-tetrahydropyridin-2-ylamine.

chloromethyl-3,4,5,6-tetrahydropyridin-2-ylamine) shown in the left-hand column has been broken down into the nine atomic-centered molecular fragments reported in the middle column. The right-hand column reports a potential set of compounds that can be associated with each fragment. The substructure is highlighted in bold. No set has been associated with singly occurring fragments. Therefore, in this case, m(i) = 6.

To mathematically express the fact that a molecular structure can be "partitioned" over the different clusters accessible to it, we introduce a membership matrix, P, consisting of N × M elements. The generic element P(i,j) indicates to what degree the i-th molecule belongs to the set, or cluster, associated with the j-th fragment. This concept of membership is analogous to that usually encountered in fuzzy logic.[33] As in fuzzy logic, we set the sum of memberships for each molecule equal to unity, so that the element P(i,j) denotes a probability. Mathematically, this condition can be expressed as

$$\sum_{j=1,M} \mathrm{P(i,j)} = 1, \forall\, \mathrm{i} \qquad (5)$$

This condition also implies that P(i,j) must be a real number in the interval [0,1]. Below, we show how interpreting P as a probability matrix can help justify the initial guess for the recursive calculation. Having introduced the matrix P gives us a simple way to represent molecules in terms of clusters. However, we also want to provide P with a counterpart, a complementary matrix that quantifies the degree to which molecules contribute to cluster populations. We therefore introduce the weight matrix W, made up of N × M elements, W(i,j). In order to find a convenient expression for W(i,j), we consider that the population of each set, C(j), is given by n(j). Summing up all individual populations, we obtain the total database population K, that is

$$\mathrm{K} = \sum_{j=1,M} \mathrm{n(j)},\ \text{for}\ \mathrm{n(j)} > 1 \qquad (6)$$

K only depends on the structure of the occupancy matrix O. In fact, K is simply the sum of all elements of the occupancy matrix, excluding singly occurring fragments. Here, we want its value to remain constant regardless of the distribution of compounds among the available clusters. Accordingly, the total sum of weights should not exceed this value:

$$\sum_{i=1,N} \sum_{\mathrm{J}=1,\mathrm{M}} \mathrm{W(i,j)} = \mathrm{K} \qquad (7)$$

In Appendix B (see the Supporting Information), we demonstrate that eq 7 can be fulfilled by setting

$$\mathrm{W(i,j)} = \mathrm{m(i)} \cdot \mathrm{P(i,j)} \qquad (8)$$

Changing the weights enables us to optimize the evaluation of each cluster, as shown below. At this point, one might ask which values must be assigned to P and W for those matrices to represent the database under investigation. Since there is no reason to distinguish among different sets, the natural choice is the following "uniform" distribution:

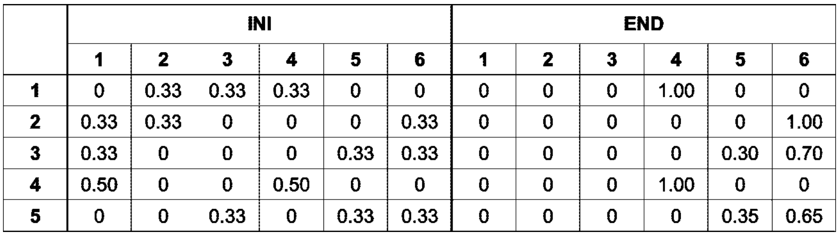
$$\mathrm{P(i,j)} = \begin{cases} \mathrm{O(i,j)/m(i)}, & \text{if } \mathrm{n(j)} > 1 \\ 0, & \text{if } \mathrm{n(j)} = 1 \end{cases} \qquad (9)$$

Combining eq 9 with eq 8, we obtain the corresponding weight matrix:

$$\mathrm{W(i,j)} = \begin{cases} \mathrm{O(i,j)}, & \text{if } \mathrm{n(j)} > 1 \\ 0, & \text{if } \mathrm{n(j)} = 1 \end{cases} \qquad (10)$$

Eqs 9 and 10 simply describe a database where the compounds are uniformly distributed over the clusters defined by a specific fragment pool. 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, according to eq 8, are then equal to unity. This configuration gives all m(i) sets the same importance in describing a molecule, regardless of their composition.

Now, suppose we want to find an approximate description of a molecule in terms of membership values: in this case, we must account for the fact that different sets describe a molecule to different extents. Sets containing molecules structurally similar to the queried compound would contribute more than others to the molecular description. Figure 1 provides a simple example: in comparison to all the other sets, the first one contains the molecule structurally most similar to the compound queried. Thus, if we wish to "structurally" describe this compound in terms of memberships, we should assign the highest membership value to the first set. Obviously, the uniform distribution does not help in this task, since it does not account for the descriptive ability of each set. Nevertheless, it can be used as a starting point for finding a distribution that best reflects each set contribution to the compound description. We might consider the uniform distribution as the initial state of a process aimed at finding the clusters better describing each compound. From



	INI						END					
	1	2	3	4	5	6	1	2	3	4	5	6
1	0	0.33	0.33	0.33	0	0	0	0	0	1.00	0	0
2	0.33	0.33	0	0	0	0.33	0	0	0	0	0	1.00
3	0.33	0	0	0	0.33	0.33	0	0	0	0	0.30	0.70
4	0.50	0	0	0.50	0	0	0	0	0	1.00	0	0
5	0	0	0.33	0	0.33	0.33	0	0	0	0	0.35	0.65

Figure 2. Examples of initial and final membership matrices.

the probabilistic point of view, starting from the uniform distribution means assuming that all m(i) sets accessible to the i-th molecule have equal probability of being selected as optimal. As mentioned above, this probability is equal to the membership value 1/m(i).

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:

$$A(i, j) = \frac{\left(\sum_{ii \in C(j)} S(i, ii) \cdot W(ii, j)\right)}{n(j) - 1} \tag{11}$$

Here, A(i, j) is the affinity between the i-th molecule and the set C(j). We chose the term “affinity” to avoid confusion with the term “similarity”, which is used in the contest of pairwise comparisons.

Of course, in order to provide the full set of values associated with the i-th molecule, eq 11 has to be worked out m(i) times. Eq 11 measures the extent to which the i-th molecule is structurally described by the remainders, n(j)-1, within the j-th set. Computing eq 11 requires knowledge of the weights W(ii,j) before the calculation. These values are given by eq 10, since we assume uniform distribution as a starting guess. Now, we need to exploit the information given by eq 11 to update the memberships in order to improve the description of each molecule. Since the higher the affinity, the better the description at a structural level, we may simply set the membership equal to a normalized affinity:

$$P(i, j) = \frac{A(i, j)}{\sum_{k=1,M} A(i, k)} \tag{12}$$

Normalization is required in order to fulfill eq 5. The new memberships obtained through eq 12 are directly proportional to the corresponding affinities. This rewards the memberships associated with the sets best describing the molecule, at the expense of all the others. The application of eqs 11 and 12 transforms the initial uniform distribution into a nonuniform one. The new distribution, with a new membership matrix, P, describes each molecule more accurately within the clustering scheme defined by the fragments used.

In order to further refine the clustering scheme, we only need to consider the new membership matrix P as input and repeat the procedure outlined above. First, we update the weight matrix using eq 8. Then we sequentially apply eqs 11 and 12 in order to obtain a new output matrix. At each iteration, the affinity function increases the membership of each molecule in the clusters best describing its structure. Molecules do not behave as individual entities during the process as molecular affinities influence each other through the weights W(ii,j) in eq 11. As the next iteration begins, the weight matrix is recomputed to update cluster memberships. Convergence is achieved when the difference between two consecutive membership matrices becomes negligible. If the process does not converge, it means the chosen clustering scheme cannot lead to any stable configuration.

Figure 2 provides an example of a recursive procedure for a simple database consisting of five molecules and six fragments. The initial—INI—and the final—END—membership matrices are shown. The INI matrix simply corresponds to the uniform distribution. The END matrix was obtained after the recursive process reached convergence and shows the different clustering scenarios available to a compound. Molecules with unitary membership value— numbers 1, 2, and 4—belong completely to exactly one cluster, while compounds 3 and 5 are partitioned over two clusters. The clusters associated with fragments 1, 2, and 3, initially populated, are now “empty”. This typical outcome indicates that these clusters had essentially no role in the final clustering scheme, and all molecules gradually “left” them.

The last effort needed to make this a tool for outlier detection is the choice of a criterion that quantifies outlying behavior. To this end, we assign each molecule an additional set, hereafter called the outlier set. Like any other set, the outlier set has a membership and a weight for each compound. We define the membership associated with the outlier set for the i-th compound as $P_o(i)$ and the corresponding weight as $W_o(i)$. The outlier set contains only the molecule it refers to and serves to quantify the extent to which the compound is *not* described by the other sets. The main idea behind the introduction of the outlier sets is to provide each compound with a sort of “reservoir”, to store poor descriptions in terms of membership. From the methodological point of view, introducing the outlier sets means adding M singly occupied sets to the cluster-based representation of the data set and requires us to modify some equations accordingly.

The normalization condition set by eq 5 should now include the outlier set for each compound:

$$P_o(i) + \sum_{j=1,M} P(i, j) = 1, \forall i \tag{13}$$

Eq 7 should now account for the addition of M sets:

$$\sum_{i=1,N} \sum_{J=1,M} W(i, j) = K + M \tag{14}$$

The presence of the outlier sets increases the number of clusters available to the i-th compound from m(i) to m(i)+1. We therefore rewrite eq 8 as

$$W(i,j) = (m(i) + 1)\cdot P(i,j) \tag{15}$$

It is possible to demonstrate that eq 15 fulfills eq 14 if

$$W_o(i) = (m(i) + 1)\cdot P_o(i) \tag{16}$$

We will show below that eq 16 does not need to be evaluated during the recursive procedure, since the affinity to the outlier sets cannot be computed. Eq 9 can also be updated by replacing "m(i)" with "m(i)+1":

$$P(i, j) = \begin{cases} O(i, j)/(m(i) + 1), & \text{for } n(j) > 1 \\ 0, & \text{for } n(j) = 0 \end{cases} \tag{17}$$

The next equation completes the uniform distribution:

$$P_o(i) = 1/(m(i) + 1) \tag{18}$$

Combining eqs 15 and 17 leaves eq 10 unchanged.

At this point, we encounter an important problem related to the evaluation of the memberships in outlier sets during the recursive calculation. In order to meet the normalization condition expressed by eq 13 we must update eq 12 in the following way

$$P(i,j) = \frac{A(i, k)}{Ao(i) + \sum_{k=1,M} A(i, k)},\ P_o(i) = \frac{Ao(i)}{Ao(i) + \sum_{k=1,M} A(i, k)} \tag{19}$$

where $A_o(i)$ is the affinity between the i-th compound and its outlier set. While the elements A(i,k) can be estimated using eq 11, $A_o(i)$ cannot, since the outlier sets are singly occupied. In order to provide eq 19 with the affinity $A_o(i)$ we set

$$A_o(i) = P_o(i) \tag{20}$$

Eq 20 avoids the evaluation of eq 16 and is used in place of eq 11 at each iteration.

Running the calculation with the outlier set has no effect on compounds completely described by the other sets. However, for a compound showing low affinity compared to all the available sets, the normalization condition (eq 13) increases the membership in the outlier set. If this process continues until the membership reaches the maximum value, the compound becomes an outlier, according to the definition given in Section 2.1. Mathematically, this definition can be stated as follows:

$$\text{If } P_o(i) = 1 \text{ then the i-th compound is an outlier} \tag{21}$$

Eq 21 can be formally applied to the final clustering scheme to identify all outliers in the database.

Figure 3 illustrates what happens when the outlier sets were added to the five molecules considered above in Figure 2. The same occurrence matrix was considered. The initial memberships changed with respect to the previous example, owing to the presence of an additional set for each compound. Like in the example above, the sets associated with fragments 1, 2, and 3 were empty at convergence. The introduction of the outlier sets did not affect the behavior of the fragments associated with empty sets. It changed, however, the contribution of the remaining fragments to compounds description.

Going from Figure 2 to Figure 3, the set associated with fragment 4 changed its role in the final clustering scheme. In the first example it was selected as the set best describing compounds 1 and 4. Adding the outlier sets, however, revealed that its contribution was not significant.

Lacking membership in nonempty sets, compounds 1 and 4 were left with no other choice than to become outliers. The necessity to fulfill the normalization condition forced the membership in the outlier set to reach the maximum value for both compounds.

In contrast to compounds 1 and 4, compound 2 was partially described by the nonempty set associated with fragment 6. This prevented the compound from becoming an outlier. At the same time, the lack of description in all other sets gave the compound a residual membership in the outlier one.

The remaining compounds, 3 and 5, are entirely described by nonempty fragments. Their membership in the outlier set approached zero during the recursive procedure.

2.3. Choice of the Fragments and the Similarity Function. In the previous section, we outlined our method in a general way, not providing fragment typology or the functional form for the similarity function. Both these important elements have to be set by the user. Here, we explain our choice for the current study.

The choice of Atomic Centered Units (ACUs) as molecular fragments was 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.[34] ACUs are small substructures, ranging from two to five atoms, made up of collecting a central parent atom and its closest neighbors. The small size and high statistical occurrence suggested they were 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[35] to the values from the occupancy matrix

$$S(i, ii) = \frac{C(i, ii)}{U(i) + U(ii) - C(i, ii)} \tag{22}$$

where

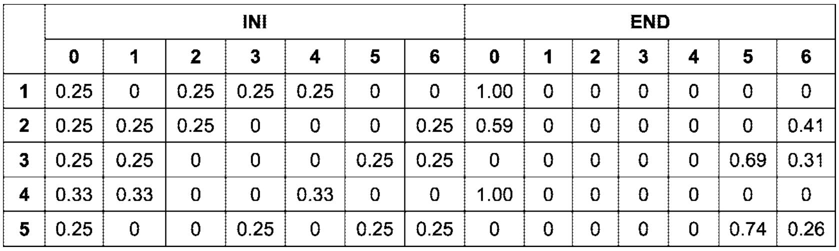
$$C(i, ii) = \sum_{j=1,M} O(i, j)\cdot O(ii, j) \tag{23}$$

$$U(i) = \sum_{j=1,M} O(i, j)\cdot O(i, j) \tag{24}$$

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

3. RESULTS

In this work, three databases were considered, each covering a different chemical composition. Duluth,[36] the largest database, consisted of 610 aliphatic and aromatic chemicals. Alcohols, esters, ketones, amines, benzene derivatives, haloalkanes, and nitro compounds were the chemical classes most represented. Demetra consisted of 282 pesticides, in more than 20 different pesticide classes.[37] These included organotins, organochlorines, organophosphates, carbamates, formamidines, terpenes, pyrethroids, phenols, spinosyns, pyrroles, pyridazinones, benzoylureas, and others. The Amines database[38] was the smallest, comprising 95



	INI							END						
	0	1	2	3	4	5	6	0	1	2	3	4	5	6
1	0.25	0	0.25	0.25	0.25	0	0	1.00	0	0	0	0	0	0
2	0.25	0.25	0.25	0	0	0	0.25	0.59	0	0	0	0	0	0.41
3	0.25	0.25	0	0	0	0.25	0.25	0	0	0	0	0	0.69	0.31
4	0.33	0.33	0	0	0.33	0	0	1.00	0	0	0	0	0	0
5	0.25	0	0	0.25	0	0.25	0.25	0	0	0	0	0	0.74	0.26

Figure 3. Examples of initial and final membership matrices after the introduction of outlier sets. The label "0" identifies the outlier sets.

Table 1. Outlier Detection Results

database	N[a]	M_{ACUs}[b]	N_{Out}[c]	$\%_{Out}$[d]
Duluth	610	597	65	10.7
Demetra	282	720	38	13.5
Amines	95	85	2	2.1

[a] Number of compounds in the data set. [b] Total number of ACUs. [c] Number of outliers. [d] Percentage of outliers with respect to total compounds (from eq 25).

aromatic and heteroaromatic amines, ranging from simple anilines to more complex phenazines, and quinolines.

Databases were processed separately. For each, the 2D chemical structures, available in MDL MOL format, were preprocessed and then broken down into ACUs, following the strategy described in a previous work.[34] The total number of ACUs obtained for each database is reported in Table 1 (M_{ACUs}). Binary ACUs occurrences were retrieved in the occupancy matrix. The pairwise similarities were computed as described in Section 2.3. Both matrices were afterward submitted to the outlier detection algorithm, schematically described in Appendix A (see the Supporting Information), and implemented in Fortran 77.

Table 1 summarizes the results of outlier detection for all databases. The outlier fraction ($\%_{Out}$) was computed using the formula

$$\%_{Out}=N_{Out}/N\cdot 100 \tag{25}$$

To begin with, we consider the number and the fraction of outliers. For Duluth, 65 compounds out of 610 ($\%_{Out}$ = 10.7%) were recognized as outliers. In Demetra, 38 outliers were detected. The corresponding fraction (13.5%) was the highest among these databases. The Amines database had the smallest number of outliers (2 out of 95 compounds) and the smallest proportion (2.1%).

At first glance, the amount of outliers for Duluth and Demetra looks higher than one would expect, thinking of outliers as merely a few "exceptions". It is, nonetheless, consistent with the definition of outlier as a compound with unitary membership in the outlier set. Considering outlyingness in terms of memberships avoids the need for thresholds. At the same time, it widens the types of outliers that can be found. In our approach, all outliers are entirely made up of combinations of two different kinds of fragments.

The first comprises fragments associated with empty sets. In Section 2.2, we have seen that, while initially all sets are populated, only a fraction contributes to the final clustering scheme. The sets lacking the ability to group structurally similar molecules are gradually "left" by all compounds during the recursive process until they become "empty" (a set is said to be "empty" when, by the end of the recursive procedure, the membership value of all its members is zero). If a compound lacks fragments associated with sets other than empty ones (like compounds 1, and 4, in Figure 3), the normalization condition increases its membership in the outlier set to compensate the lack of description. Since the contribution of empty sets to compounds description is null at convergence, the compound becomes an outlier.

It is important to note that, such a "compensation" mechanism does not account for the position of the compound in the descriptor space. The recursive procedure does not assess outlyingness on the basis of the distance between compounds; it only quantifies the descriptive ability of each fragment taking into account the positions of all compounds.

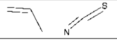
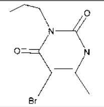
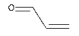
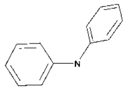
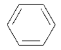
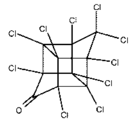
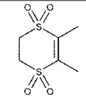
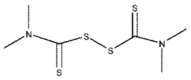
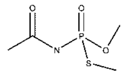
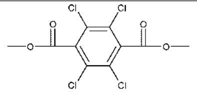
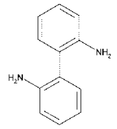
The descriptive ability of the fragments determines the outlier character. A compound can be well described in terms of the statistical occurrences of its fragments but, nonetheless, recognized as an outlier because these fragments are associated with empty sets and, therefore, lack descriptive ability. This explains why, in our approach, an outlier is not necessarily a compound structurally different from all the others. We shall return to this important point in the Discussion.

Another class of fragments that, like the previous one, do not contribute to compounds description and can be found in outliers is that of SOFs. SOFs are unique structural features. Although they do not directly participate in the recursive process, their presence might indirectly affect outlier detection. SOFs are explicitly taken into account in computing the pairwise similarities. Their presence in a compound reduces its similarity to all the other molecules, increasing its membership in the outlier set. At the same time, SOFs occurrence reduces the number of sets available for clustering. From a statistical point of view, this means less chance of finding clusters containing similar compounds. For both reasons, a compound containing SOFs is, in principle, more likely than others to become an outlier.

So far, we have identified the structural features that characterize all outliers. An outlier can be made up of any combination of such features. This considerably diversifies the outliers in terms of structural resemblance with other compounds. Outliers with high SOFs content look like standalone compounds in the chemical space, while those characterized by low SOFs content may share considerable portions of their structures with other compounds. A method based on the similarity alone would detect the first class of outliers, but it would probably miss the second one. The fact that our approach deals with both types of outliers at once is responsible for the high fraction of outliers found in Duluth and Demetra.

Given this consideration, we may examine some examples of outliers extracted from all the data sets under investigation.

Table 2. Examples of Outliers

ID[a]	Dataset[b]	Name[c]	Compound[d]	m_{ACUs}[e]	m_{SOFs}[f]	m_{EMPs}[g]
1	Duluth	Allyl isothiocianate 57-06-7		6	4	2
2	Duluth	Bromacil 314409		13	5	8
3	Duluth	Acrolein 107-02-8		4	1	3
4	Duluth	Diphenylamine 122394		4	1	3
5	Duluth	Benzene 71432		1	0	1
6	Demetra	Chlordecone 143-50-0		6	3	3
7	Demetra	Dimethepin 55290-64-7		5	3	2
8	Demetra	Thiram 137-26-8		5	2	3
9	Demetra	Acephate 30560-19-1		10	4	6
10	Demetra	Dacthal 1861-32-1		7	0	7
11	Amines	2,2′- Diaminobiphenyl 1454-80-4		5	0	5

[a] Compound identifier. [b] Source. [c] Compound common name and CAS number. [d] Compound sketch. [e] Number of ACUs. [f] Number of SOFs. [g] Number of fragments associated with empty sets.

These compounds, reported in Table 2, were selected as the representatives of different molecular situations.

Compound **1**, allyl isothiocianate, is made up of six different ACUs, four occurring singly. This compound does not structurally resemble any other in the data set. The SOFs reduced its membership in the remaining two sets (which became empty at convergence), increasing the membership in the outlier one. All compounds with high SOFs content quickly become outliers during the recursive procedure.

Compound **2**, bromacil, is also a compound with no significant analogs in the data set. Its structure is characterized by 5 SOFs. The remaining fragments had no role in the recursive procedure, meaning that none of them was able to cluster structurally similar compounds.

Compound **3**, acrolein, is characterized by low SOFs content. Nonetheless, its structure is unique in the data set. The molecule is made up of four ACUs, one occurring singly. The remaining ACUs define sets containing structurally dissimilar molecules. Acrolein is likely to become an outlier early during the recursive procedure, on account of its very low membership values.

Compound **4**, diphenylamine, is a typical example of a compound that is made by a low number of ACUs (4) that are mostly members of empty set (3). The low number of ACUs decreases the chance of finding representative fragments.

Compound **5**, benzene, is made up of only one ACU that occurs very frequently in the database. This molecule is well represented in terms of the statistical occurrences of ACUs. However, this ACU, does not identify any specific group of compounds and quickly loses importance during the recursive procedure. This typically happens to fragments widely occurring in the database, such as the ACUs from the breakdown of aromatic rings, or aliphatic chains.

The next five examples of outliers selected for discussion come from the Demetra data set. Compound **6**, chlordecone, was recognized as an outlier because of its unique chemical structure. The same applies to compound **7**, dimethepin. Also this compound is characterized by high SOFs content and has no analogs in the data set.

As the SOFs content decreases, empty sets start playing a leading role in outlier detection. Compounds **8**, thiram, and **9**, acephate, are better represented than previous compounds (only 40% of SOFs), considering the occurrence of ACUs individually. But, the combination of some SOFs with some ACUs that are widespread in the data set has the effect of precluding the selection of a suitable cluster. The result is their insertion into the outlier set.

Finally, compound **10**, dacthal (from Demetra data set), and compound **11**, 2,2′-diaminobiphenyl (from the Amines data set), are entirely made up of very common fragments (no SOFs). Again, what makes them "different" from the other compounds is not the presence but the absence of significant structural features.

4. DISCUSSION

Two preliminary points should be clarified before beginning the discussion. First, we want to corroborate the use of the term "outlier" in our approach. As mentioned in the Introduction, this term usually identifies standalone compounds in the chemical space defined by the set of descriptors used. Here, this term still identifies compounds that are positioned out of a given chemical space. The difference is, however, that in our approach this space is not defined by all descriptors but only by those associated with nonempty sets. We can say that the recursive procedure reduces the original chemical space to a subspace where outlier detection is performed. According to this picture, a compound entirely described by fragments associated with empty sets is an outlier since it is located outside the space defined by the fragments that contribute to compounds description.

A second aspect concerns the choice of the fragment-based representation. In this regard, it is important to note that our method makes no attempt to assess the validity of a molecular representation. That is, the recursive procedure only evaluates the contribution of each set to compounds description. It does not assess whether the fragments are representative of the compounds under investigation. From the methodological point of view, there is no difference between a good and a bad representation: both can be successfully analyzed by our method.

This does not mean that the quality of the representation does not affect that of the result. Among all possible representations, we may identify the suitable ones on the basis of three attributes. First of all, the fragment pool must be representative of all compounds in the database; incomplete or partial description of the chemical structures would bias the detection process and must be avoided. Then, suitability also demands statistically significant fragments. Finally, the “portability” of a representation must be considered.

The way fragments are generated is a key factor in meeting these requirements. Since the use of a collection of predefined fragments, such as functional groups, might lead to an incomplete description of the molecular structures, and often to a lack of portability, a better option would be to extract the chemical features directly from the database. Automatic fragment generation guarantees full coverage of the molecular structures, and high portability, since the same rule can be applied across different databases. Statistical significance can also be guaranteed simply by tuning fragment specificity to the desired level.

Given these recommendations, we may reasonably ask what would happen to the occurrence of outliers in passing from one fragment-based description to another. Assuming the chosen fragment-based representation fulfills the above requirements, we may expect that increasing fragment specificity would increase the number of outliers found. Alternatively, reducing the fragment size and specificity means fewer outliers are found. To test this, we ran an experiment on Demetra, using DFGs[39] (Diatomic FraGments) instead of ACUs. DFGs are obtained considering two neighbor atoms and the bond connecting them. DFGs are less specific than ACUs; 186 DFGs were needed to describe the entire data set, compared to 720 ACUs. Using the same similarity function (see eq 16), we detected 12 outliers instead of 38. Nine outliers were common to both descriptions; the remaining compounds changed their status, from outlier to nonoutlier, or vice versa.

Over the years, several criteria for assessing the correctness of outlier selection have been used, from simple visual inspection to more sophisticated ones.[40–42] Regardless of their efficacy, however, these criteria are subjective: they incorporate the user’s perception in the discovery process. It is therefore not possible to establish whether an outlier selection is correct or not. The existence of conflicting assignments is therefore compatible with the existence of different representations. There is no need to establish which description is the most reliable, because all are reliable. The preference for a specific description can only be dictated by the application.

In the Introduction, we highlighted the importance of the descriptors in the search for outliers. It should be clear that a compound is an outlier in a given descriptors space only in relation to the competence that the descriptors have in representing that compound. As a consequence, what we really identify is not the outlier compound but the failure of the descriptors to represent that compound. The novel opportunity that we introduced in the present work concerns the assignment of a “fail” label to descriptors that are not unique in the data set, vice versa they are so common to become source of uncertainty. In this way, a particular compound can be defined as an outlier on the basis of the very low importance of its descriptors in giving significant information.

Outlier detection in a training set can be helpful in order to recognize the descriptors that are less informative for the model. This result is important because it permits a better perception of the model quality. A consequence is that we can decide to use all the available data, but we know that some of them can be a source of uncertainty. In addition, outlier detection greatly helps evaluate the prediction reliability of a given model for an unknown molecule. In this case, the query compound is added to the training set, and the resulting database is then inspected for outliers. The main idea is to establish in advance whether the model is suitable for the compound queried rather than to calculate the prediction accuracy. This strategy is very useful indeed in predicting chemical properties of compounds for which no experimental data are available.

Different scenarios can be observed, that may lead to unreliable estimates of chemical properties. The most typical and frequent scenario is when the query compound has no structural analogues in the training set. In this case, the lack of representation is the factor responsible for poor prediction reliability. A second scenario is when the query compound is structurally represented by descriptors present in many molecules that do not share any significant descriptor. The resulting prediction is therefore “averaged” over a multitude of potentially different chemical behaviors; it is reasonable to expect more accurate prediction as the number of different behaviors in each group decreases. The last scenario is the most challenging and occurs when only a few structurally diverse analogues are available to describe the compound queried. In this case, the query molecule might be fully described by its analogues and therefore not identified as an outlier. The description is not only complete but also fragmented, since each analogue only captures a limited portion of the molecular structure. Such a representation reduces the query molecule to a collection of isolated chemical domains, whose interactions and contributions to the overall property value are not usually known. Without this information, the model response cannot be adequately controlled, and the resulting estimate should not be trusted. Our approach does not solve this problem. However, as far as we know, it is not solved by the other approaches either, leaving room for developments.

We do not apply our method to data sets used for drug discovery. Nevertheless, we can think that the identification of structural outliers can be useful also in this field. Our method allows for the location of compounds that only contain critical fragments; i.e. these compounds potentially behave differently with respect to the other members of the data set and, therefore, need special attention.

The Duluth and Amine databases are well-known and often used databases. In the literature,[39,43–48] there are many models that predict their compound activity with good accuracy. However, as far as we know, this is the first time when an *a priori* study on outlier existence is done. All existing papers only discuss the existence of outlier compounds on the basis of the prediction accuracy. It should be clear that these results cannot be directly compared with ours. The occurrence of bad predicted compound activities can be the consequence of several causes; e.g. an inadequate model or a wrong experimental determination. In contrast, the *a priori* selection of problematic descriptors is completely independent of the prediction model; it only helps in the removal of scarcely useful data. It should be clear that the removal of some too generic descriptors does not imply the impossibility to predict the activity of a compound; it only prevents the assignment of wrong weights to the fragments forming the compound.

CONCLUSIONS

Detection of outliers is an important and challenging issue; it has several applications and can give interesting hints for understanding chemical behaviors better. However, as often happens when exploring elusive concepts, an additional complication is the lack of a universally accepted definition of the term we are examining. Outlyingness is a property of an object that is strictly related to the field of application and to the detection method. We have made an effort here to give a new and general definition of this property, connecting outlyingness to the properties of a chemical space and also giving an exact description of the space in terms of attributes. If we accept the idea that a chemical space is a dynamic concept depending on the attributes used for its description, we are in a better position to describe its outlyingness or, better, what an object in the space needs to be defined as an outlier. We would like to highlight this point because it is the core of the work. The explicit reference to the object membership in a set has the consequence of making a clear definition of an outlier: the absence of unambiguous membership to any set of the space. This definition is so clear that it also permits the identification of a new variety of outliers, those that are “spread” over too many sets.

We also applied the method for outlier detection to some examples. The results are indicative of the performance of our approach. If we have a good data set description and an unbiased search procedure, we can confidently select the compounds that are outliers in a chemical space. It is then up to the user to choose the description and detection function best fitting the needs. Nevertheless, our description and function generate the expected result.

The more general problem of outlier validation remains. Currently, there is no single answer; we can only suggest minimizing the parameters to be set and accurate application of well-defined procedures in order to limit unclear results.

ACKNOWLEDGMENT

The authors gratefully acknowledge the EU project Chemomentum (contract number: IST-5-033437) for having partially supported and funded the current study.

Supporting Information Available: Implementation of the algorithm (Appendix A) and derivation of equation 8 (Appendix B). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Hodge, J. V.; Austin, J. A Survey of Outlier Detection Methodologies. *Artif. Intell. Rev.* **2004**, *22*, 85–126.

(2) Guha, R.; Dutta, D.; Jurs, C. P.; Ting, C. R-NN Curves: An Intuitive Approach to Outlier Detection Using a Distance Based Method. *J. Chem. Inf. Model.* **2006**, *46*, 1713–1722.

(3) Leach, R. A.; Hann, H. M. The in silico world of virtual libraries. *Drug Discovery Today* **2000**, *5*, 326–336.

(4) Jorgensen, W. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813–1818.

(5) Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7*, 1047–1055.

(6) Guha, R.; Dutta, D.; Jurs, C. P.; Ting, C. Scalable Partitioning and Exploration of Chemical Spaces Using Geometric Hashing. *J. Chem. Inf. Model.* **2006**, *46*, 321–333.

(7) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; Van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. The Report and Recommendations of ECVAM Workshop 52. *ATLA* **2005**, *33*, 155–173.

(8) He, L.; Jurs, P. C. Assessing the reliability of a QSAR model’s predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503–523.

(9) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. Review of methods for assessing the applicability domains of SARs and QSARs. Final report to the Joint Research Centre (Contract No. ECVA-CCR. 496575-Z). Part 1: Review of statistical methods for QSAR AD estimation by the training set, 2005. European Chemicals Bureau Web Site. http://ecb.jrc.it/documentation/ (accessed Feb 8, 2008).

(10) Furusjö, E.; Svenson, A.; Rahmberg, M.; Andersson, M. The importance of outlier detection and training set selection for reliable environmental QSAR predictions. *Chemosphere* **2006**, *63*, 99–108.

(11) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.

(12) Gombar, V. K.; Enslein, K. Assessment of n-octanol/water partition coefficient: when is the assessment reliable. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1127–1134.

(13) Xu, Y. J.; Gao, H. Dimension related distance and its application in QSAR/QSPR model error estimation. *QSAR Comb. Sci.* **2003**, *22*, 422–429.

(14) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *J. Mol. Struct. (Theochem)* **2003**, *622*, 39–51.

(15) Verma, R. P.; Hansch, C. An approach toward the problem of outliers in QSAR. *Bioorg. Med. Chem.* **2005**, *13*, 4597–4621.

(16) Cronin, M. T. D.; Sinks, G. D.; Schultz, T. W. In *Forecasting the Environmental Fate and Effects of Chemicals*, 1st ed.; Rainbow, P. S., Hopkins, S. P., Crane, M., Eds.; Wiley: Chichester, England, 2001; pp 111−113.

(17) Kirchner, L. A.; Moody, R. P.; Doyle, E.; Bose, B.; Jeffry, J.; Chu, I. The prediction of skin permeability by using physicochemical data. *ATLA* **1997**, *25*, 359–370.

(18) Commission of the European Communities Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency and amending Directive 1999/45/EC and Regulation (EC) {on Persistent Organic Pollutants}, 2003. European Union law Web Site. http://europa.eu/eur-lex/en/com/pdf/2003/com2003_0644en.html (accessed May 14, 2008).

(19) Golbraikh, A. Molecular data set Diversity Indices and Their Applications to Comparison of Chemical Databases and QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 414–425.

(20) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. P.; Ivaschenko, A. A. New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249–258.

(21) Jorgensen, W. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods. *Science* **2004**, *303*, 1813–1818.

(22) Agrafiotis, D. K. A Constant Time Algorithm for Estimating the Diversity of Large Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 156–167.

(23) Godden, J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. Median Partitioning: A Novel Method for the Selection of Representative Subsets from Large Compound Pools. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 885–893.

(24) Schnur, D. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36–45.

(25) Vracko, M. Kohonen Artificial Neural Network and CP NN. *Curr. Comput.-Aided Drug Des.* **2005**, *1* (1), 73–78.

(26) Guha, R.; Serra, J., R.; Jurs, P. C. Generation of QSAR sets with a self-organizing map. *J. Mol. Graphics Modell.* **2004**, *23*, 1–14.

(27) Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I.; Aptula, A. O. Structure-toxicity relationships for aliphatic chemicals evaluated with Tetrahymena pyriformis. *Chem. Res. Toxicol.* **2002**, *15*, 1602–1609.

(28) Gramatica, P.; Pilutti, P.; Papa, E. Predicting the NO_3 radical tropospheric degradability of organic pollutants by theoretical molecular descriptors. *Atmos. Environ.* **2003**, *37*, 3115–3124.

(29) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. Methods for reliability, uncertainty assessment, and applicability evaluations of classification and regression based QSARs. *Environ. Health Perspect.* **2003**, *111*, 1361–1375.

(30) Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **2004**, *112*, 1249–1254.

(31) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. *Altern. Lab. Anim.* **2005**, *33*, 445–459.

(32) Plant, C.; Böhm, C.; Tilg, B.; Baumgartner, C. Enhancing instance-based classification with local density: a new algorithm for classifying unbalanced biomedical data. *Bioinformatics* **2006**, *22*, 981–988.

(33) Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*; Plenum Press: New York, 1981.

(34) Casalegno, M.; Sello, G.; Benfenati, E. Top-Priority Fragment QSAR Approach in Predicting Pesticide Aquatic Toxicity. *Chem. Res. Toxicol.* **2006**, *19*, 1533–1539.

(35) 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.

(36) ECOTOX Database. US-EPA. http://cfpub.epa.gov/ecotox (accessed February 20, 2007).

(37) Benfenati, E.; Casalegno, M.; Cotterill, J.; Price, N.; Spreafico, M.; Toropov, A. Characterization of Chemical Structures In *Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes*, 1st ed.; Benfenati, E., Ed.; Elsevier: Amsterdam, The Netherlands, 2007; Chapter 3, pp 102−107.

(38) Sugimura, T. Overview of carcinogenic heterocyclic amines. *Mutat. Res.* **1997**, *376*, 211–219.

(39) Casalegno, M.; Benfenati, E.; Sello, G. An Automated Group Contribution Method in Predicting Aquatic Toxicity: The Diatomic Fragment Approach. *Chem. Res. Toxicol.* **2005**, *18*, 740–746.

(40) Yu, D.; Sheikholeslami, G.; Zhang, A. Finding Outliers in Very Large Datasets. *Knowl. Inf. Syst.* **2002**, *4*, 387–412.

(41) Williams, G. ; Baxter, R. ; He, H. ; Hawkins, S. ; Gu, L. In *A comparative study of RNN for outlier detection in data mining*, Proceedings of the Second IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, 2002; IEEE Computer Society Press: Los Alamitos, 2002; pp 709−712.

(42) Huang, T.; Qin, X.; Chen, C.; Wang, Q. Density-Based Spatial Outliers Detecting In *Computational Science − ICCS 2005*; Sunderam, V. S. et al. Eds.; Springer: Berlin, Heidelberg, 2005; Vol. LNCS 3514, pp 979−986.

(43) Casalegno, M.; Benfenati, E.; Sello, G. Application of a Fragment−based Model to the Prediction of the Genotoxicity of Aromatic Amines. *Int. Elect. J. Mol. Des.* **2006**, *5*, 431–446.

(44) Maran, U.; Karelson, M.; Katritzky, A. R. A Comprehensive QSAR Treatment of the Genotoxicity of Heteroaromatic and Aromatic Amines. *Quant. Struct.-Act. Relat.* **1999**, *18*, 3–10.

(45) Cash, G. G.; Anderson, B.; Mayo, K.; Bogaczyk, S.; Tunkel, J. Predicting genotoxicity of aromatic and eteroaromatic amines using electrotopological state indices. *Mutat. Res.* **2005**, *585*, 170–83.

(46) Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. *Chem. Rev.* **2000**, *100*, 3696–3714.

(47) Martin, T. M.; Young, D. M. Prediction of the Acute Toxicity (96-h LC_{50}) of organic compounds to the Fathead Minnow (Pimephales promelas) using a group contribution method. *Chem. Res. Toxicol.* **2001**, *14*, 1378–1385.

(48) Papa, E.; Villa, F.; Gramatica, P. Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow). *J. Chem. Inf. Model.* **2005**, *45* (5), 1256–1266.

CI7004065