Supplementary Table S1

Correspondence between qualitative and quantitative solubility scales used in this study

Classification	Solubility
Soluble or miscible	> 1 gr / 100 ml
Slightly soluble	0.05 - 1 g / 100 ml
Sparingly soluble	0.005 - 0.05 g / 100 ml
Insoluble	< 0.005 g / 100 ml

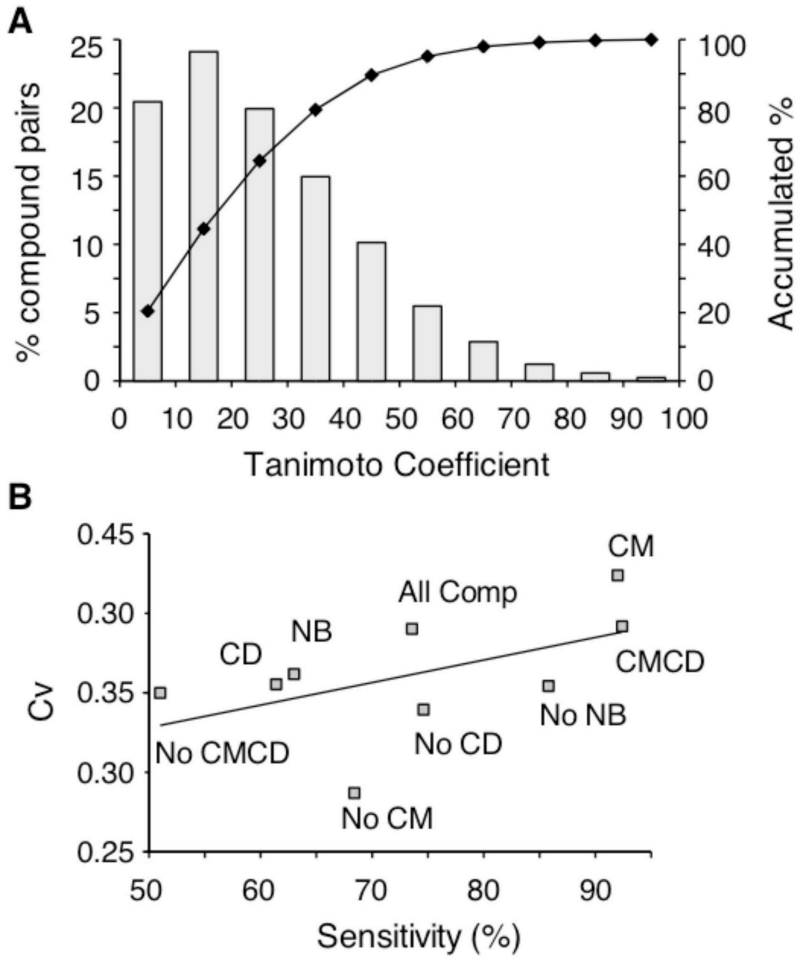
Supplementary Table S2

Prediction of environmental fates for compounds that belong to the Annex I-S, HPVC-S and LPVC-S lists

Environmental fate[a]	Annex I-S	HPVC-S	LPVC-S
CM / No CM	57/ 36	61/ 32	62/ 33
NB / No NB	51/ 43	42/ 50	49/ 45
CD / No CD	24/ 70	19/ 73	21/ 73
CMCD / No CMCD	68/ 26	73 / 19	71/ 24

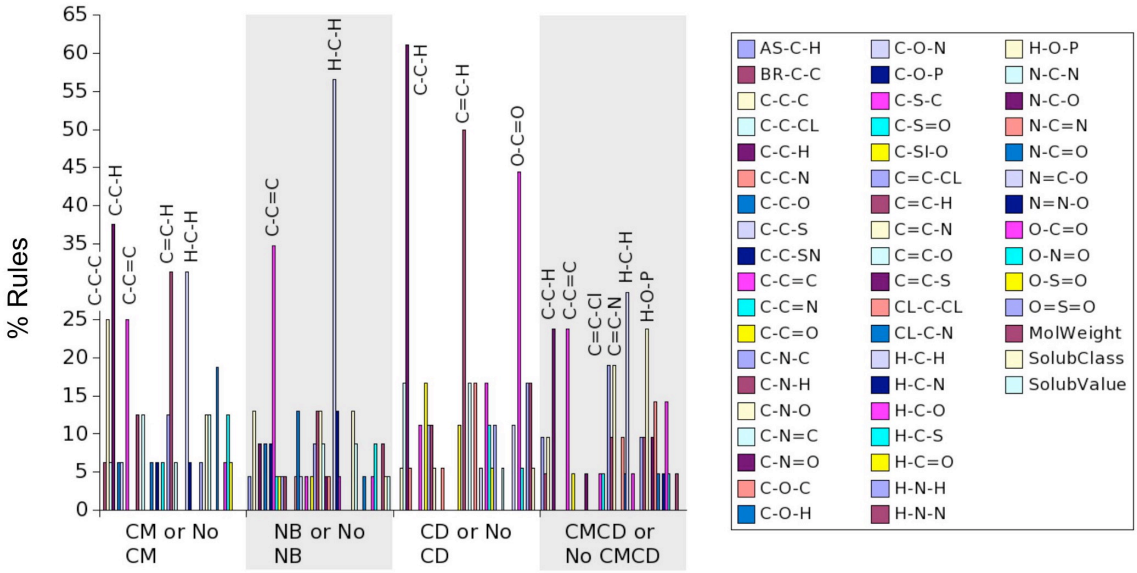
[a] The figures indicate the percentage of compounds within each of the lists that are predicted to belong to any of the classes. CM, central metabolism path compounds; NB, non-biodegradable path compounds; CD, carbon dioxide path compounds; CMCD, central metabolism and carbon dioxide path compounds.

Supplementary Figure S1. Structural similarity of the chemical compounds under study.



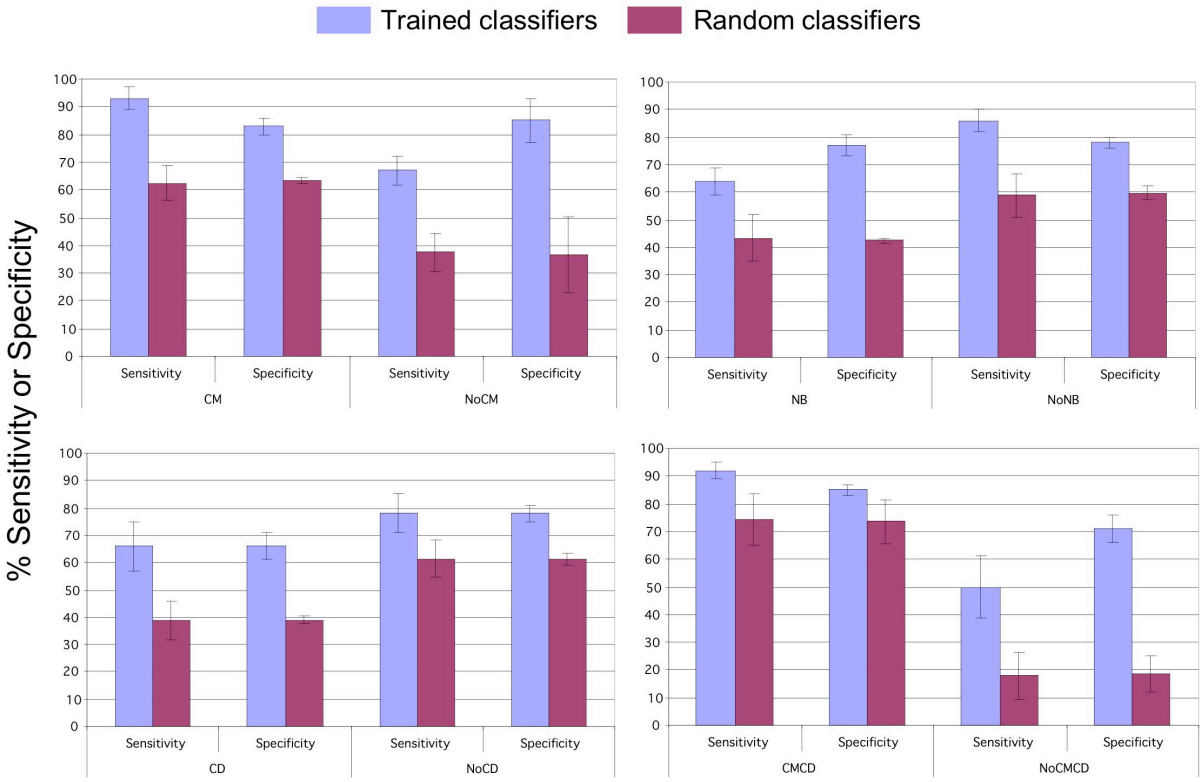
A. Distribution of Tanimoto association coefficient (τ) values for all pairs of chemical compound. Absolute frequencies are shown as bars referred to the Y axis to the left. The accumulated frequencies are shown as a curve referred to the Y axis to the right. **B.** Scatterplot representing the average clustering coefficient (*Cv*), calculated for a τ threshold of 80, versus prediction sensitivity, for compounds in the whole set of chemicals (All Comp) and in the various classes defined according to their environmental fate. The relationship between the two variables is defined by the regression line $Cv = 0.00141887 \times \text{sensitivity} + 0.25711478$, which has an associated regression coefficient $r = 0.53$. Note that about 90% of the pairwise comparisons through the complete collection of compounds were characterized by a $\tau \leq 50$ whereas only 1% of the pairs had a $\tau \geq 80$. Furthermore, the distribution of τ values was not significantly different when calculated for the various groups of compounds (CM, NB, etc., not shown) indicating that all groups are equally diverse. However, some differences among the groups were observed when the average clustering coefficient was calculated for each of them. Using a τ threshold of 80%, the average clustering coefficient for the whole collection of compounds was 0.39, while for the various groups it ranged from 0.42 to 0.286 (see **B**). The group with the highest value was CM, indicating that it contains more clusters of similar compounds that any of the other groups. On the contrary, the groups corresponding to the negated classes (No CM, No NB, etc) had the lowest clustering coefficients, indicating that are more structurally diverse than the rest. When the τ threshold was lowered to 50%, the differences in average clustering coefficient were less pronounced, ranging from 0.647 for CM to 0.574 for No CM (not shown).

Supplementary Figure S2. Frequencies of attribute usage in the rules that conform the final four classifiers.



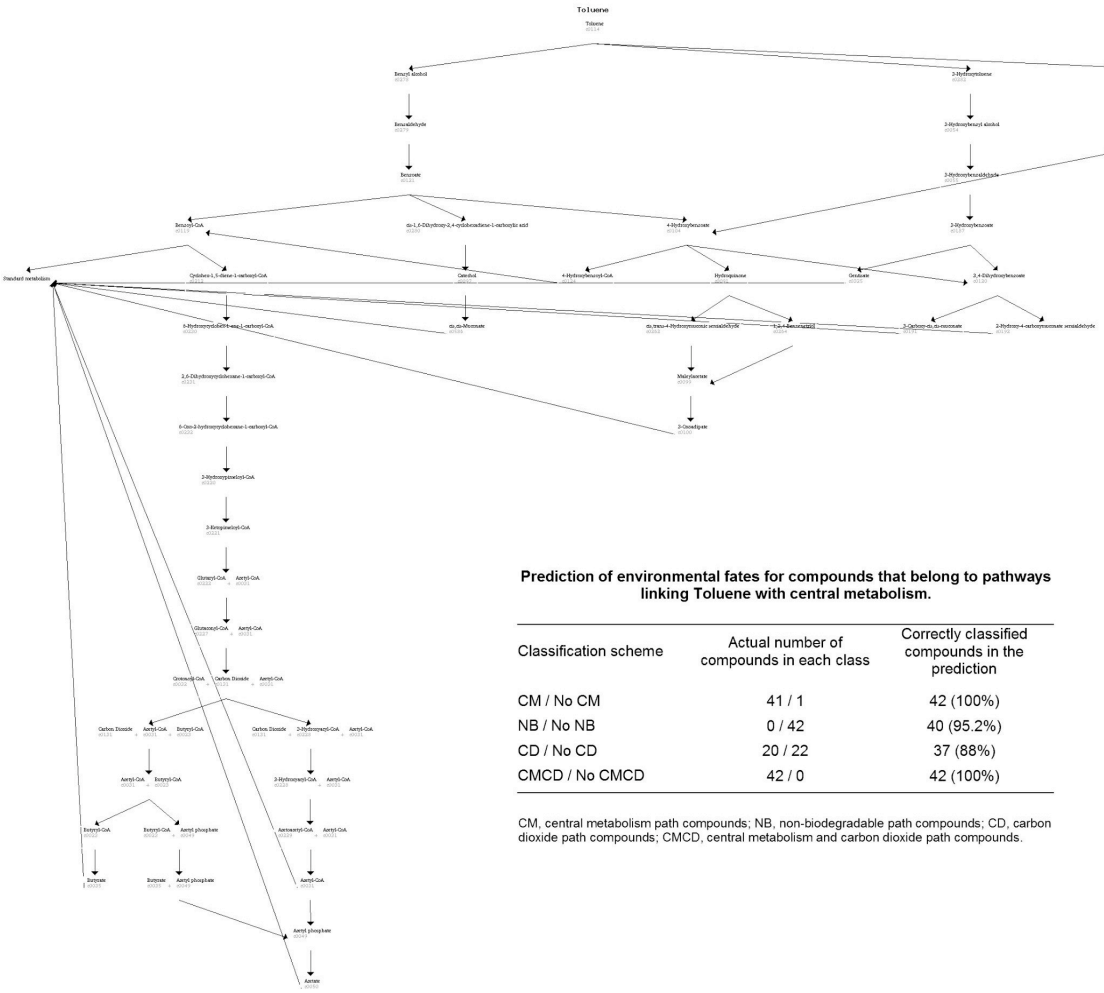
The classifiers were analyzed to determine the percentage of rules that include each of the attributes. Out of the original 152 attributes used to describe the chemical compounds, only the 52 that appear in the list were included as part of the propositional rules of all classifiers. The attributes labeled with their name, are those that appear in more than 18% of the rules.

Supplementary Figure S3. Comparison of the predictive performance of the classifiers obtained with real and randomized data.



Five-fold cross validation tests were conducted, for each of the considered classification schemes, using both the original classifiers and the equivalent random counterparts. Random classifiers assign compounds to the various classes with a probability proportional to the population of each class. The bars represent the averaged sensitivity (percentage of compounds correctly classified as belonging to a specific class, relative to the total number of cases of that particular class) or specificity (percentage of compounds correctly classified as belonging to a specific class, relative to the total number of predictions for that particular class) of the five iterations of the cross validation experiment. The error bars represent the standard deviation.

Supplementary Figure S4. Pathways for toluene biodegradation and predictive analysis of metabolites involved.



Prediction of environmental fates for compounds that belong to pathways linking Toluene with central metabolism.

Classification scheme	Actual number of compounds in each class	Correctly classified compounds in the prediction
CM / No CM	41 / 1	42 (100%)
NB / No NB	0 / 42	40 (95.2%)
CD / No CD	20 / 22	37 (88%)
CMCD / No CMCD	42 / 0	42 (100%)

CM, central metabolism path compounds; NB, non-biodegradable path compounds; CD, carbon dioxide path compounds; CMCD, central metabolism and carbon dioxide path compounds.

The network of all possible routes for microbial mineralization of toluene is show as generated with MetaRouter (Pazos *et al.*, 2005). The inserted Table shows the measure of the capacity of the Biodegradation Prediction System described in this work when applied to the whole of metabolites involved. A detailed blowup of the Toluene degradation network of the figure can be downloaded in:

http://www.pdg.cnb.uam.es/BDPSERVER/Paper/TolPath_Fig_Table.png