TyPol - a new methodology for organic compounds clustering based on their molecular characteristics and environmental behavior

Rémi Servien, Laure Mamy, Ziang Li, Virginie Rossard, Eric Latrille, Fabienne Bessac, Dominique Patureau, Pierre Benoit

TyPol - a new methodology for organic compounds clustering based on their molecular characteristics and environmental behavior

Rémi Servien[a,b,*], Laure Mamy[c], Ziang Li[d], Virginie Rossard[b], Eric Latrille[b], Fabienne Bessac[e,f,g], Dominique Patureau[b], Pierre Benoit[d]

[a] *INRA, Université de Toulouse, UMR 1331 Toxalim, Research Centre in Food Toxicology, F-31027 Toulouse, France (present address)*

[b] *INRA, UR 050, Laboratoire de Biotechnologie de l'Environnement, Avenue des Etangs, F-11100 Narbonne, France*

[c] *INRA, UR 251 PESSAC, Route de St Cyr, F-78026 Versailles, France*

[d] *UMR 1091 INRA-AgroParisTech, Environnement et Grandes Cultures, F-78850 Thiverval-Grignon, France*

[e] *Université de Toulouse, INPT, Ecole d'Ingénieurs de Purpan, Equipe DINA, 75 voie du TOEC, BP 57611, F-31076 Toulouse Cedex 03, France*

[f] *Université de Toulouse, UPS, IRSAMC, Laboratoire de Chimie et Physique Quantiques, 118 route de Narbonne, F-31062 Toulouse, France*

[g] *CNRS (UMR 5626), F-31062 Toulouse, France*

[*] Corresponding author. Address: INRA, Université de Toulouse, UMR 1331 Toxalim, Research Centre in Food Toxicology, F-31027 Toulouse, France. Tel.: +33 (0)5 61 19 32 82; fax: +33 (0)5 61 19 39 17.

E-mail address: remi.servien@toulouse.inra.fr (R. Servien)

HIGHLIGHTS

- An innovative methodology, TyPol, was developed to classify organic compounds
- The classification is based on environmental behavior and molecular descriptors
- It relies on partial least squares analysis and hierarchical clustering
- The degradation products of organic compounds are considered
- The environmental behavior of a "new" compound can be assessed from its affiliation to one cluster

ABSTRACT

Following legislation, the assessment of the environmental risks of 30 000 to 100 000 chemical substances is required for their registration dossiers. However, their behavior in the environment and their transfer to environmental components such as water or atmosphere are studied for only a very small proportion of the chemical in laboratory tests or monitoring studies because it is time-consuming and/or cost prohibitive. Therefore, the objective of this work was to develop a new methodology, TyPol, to classify organic compounds, and their degradation products, according to both their behavior in the environment and their molecular properties. The strategy relies on partial least squares analysis and hierarchical clustering. The calculation of molecular descriptors is based on an in silico approach, and the environmental endpoints (i.e. environmental parameters) are extracted from several available databases and literature. The classification of 215 organic compounds inputted in TyPol for this proof-of-concept study showed that the combination of some specific molecular descriptors could be related to a particular behavior in the environment. TyPol also provided an analysis of similarities (or dissimilarities) between organic compounds and their degradation products. Among the 24

degradation products that were inputted, 58% were found in the same cluster as their parents. The robustness of the method was tested and shown to be good. TyPol could help to predict the environmental behavior of a "new" compound (parent compound or degradation product) from its affiliation to one cluster, but also to select representative substances from a large data set in order to answer some specific questions regarding their behavior in the environment.

Keywords:
Pesticides
Degradation products
Clustering
Molecular modeling
Environmental fate
Partial least squares

1. Introduction

New legislations such as the REACH (Registration, Evaluation, Authorization and restriction of CHemicals) regulation in the EU will require that manufacturers of substances and formulators register and provide prescribed eco/toxicological data for substances with volume higher than one metric ton per year. It is estimated that about 30 000 existing substances have to be registered by 2018 by member states (Ahlers et al., 2008). The needed information has to be equivalent to the standard information requirement and adequate to draw overall conclusions with respect to the regulatory endpoints classification and labeling. Beyond specific regulatory needs, the same questions concern chemical substances that are potentially present in the

environment and that originate from various sources. According to authors, from 30 000 to 100 000 chemical substances may be concerned by environmental risks assessment (Muir and Howard, 2006). However, their behavior in the environment and their transfer to environmental components such as water or atmosphere are studied for only a very small proportion of the chemical in laboratory tests or monitoring studies because it is time-consuming and/or cost prohibitive. Consequently, a high number of different in silico approaches have been developed to estimate the behavior of organic compounds in the environment. The most used in silico approaches, that are also the most simple, are based on QSAR (Quantitative Structure Activity Relationship). QSARs allow the estimation of one or several compound properties (such as sorption by soils and sediments, biodegradation, bioconcentration factor or biological activities) from some other properties such as structural molecular properties (number of atoms in the molecule, molecular surface, dipole moment, energy of orbitals…), water solubility or octanol-water partition coefficient (e.g. OECD, 1993a; Raymond et al., 2001; Worrall, 2001; Eriksson et al., 2002; Pavan et al., 2008). Other approaches aim, for example, at ranking organic compounds knowing the values of several of their properties such as partitioning, persistence or bioaccumulation. Compounds that have undesirable properties can be considered for management, regulation, or even global bans on production and use (Mackay et al., 2001; Walker and Carlsen, 2002). Numerical models represent the most complex approaches as they allow overall assessment of the fate of organic compounds in the soil, water and air, and as they take into account the physico-chemical properties of the compounds and the agro-pedo-climatic conditions (e.g. Carsel et al., 1998; Jarvis and Larsbo, 2012). However, they require a lot of input data. Therefore, approaches able to classify compounds according to their environmental behavior or eco/toxicological effects will help both regulators and scientists facing the problem

of the constant increase in the diversity and in the number of the chemical substances which will be concerned by environmental risk assessment.

The objective of this work was thus to develop a new simple approach, TyPol (Typology of Pollutants), to classify organic compounds and their degradation products according to both their behavior in the environment and their structural molecular properties.

TyPol, is based on statistical analyses combining several environmental endpoints (i.e. environmental parameters such as sorption coefficient, degradation half-life or bioconcentration factor), and structural molecular descriptors (number of atoms in the molecule, molecular surface, dipole moment, energy of orbitals…) (Fig. 1). The calculation of molecular descriptors is based on in silico approach, and the environmental parameters are extracted from available databases and from literature. Knowing the values of several relevant structural molecular descriptors, TyPol will allow the classification of one organic compound of interest (parent or degradation product) in a group of compounds having similar values of molecular descriptors and environmental parameters, and potentially a similar environmental behavior.

The choice of the statistical method involved in TyPol is crucial for the reliability of the clustering. Principal components analysis (PCA) is often used in multivariate chemical characterizations to determine linearly uncorrelated variables that summarize the information contained in variables (Jackson, 1991; Snarey et al., 1997; Harju et al., 2002; Eriksson et al., 2006). These uncorrelated variables can also be used as an excellent basis to select a representative set of chemicals using clustering methods. Various clustering techniques have been employed in chemical mapping such as strategies based on PCA and hierarchical clustering for selecting dissimilar organic substances (Rännar and Anderson, 2010) or ranking non-ionic organic pesticides (Gramatica et al., 2004), bayesian classifiers for chemical toxicity predictions

(Mishra et al., 2011), network clustering (Saito et al., 2010), PCA-based method (Rännar and Anderson, 2011) or other statistical tools (Vogt and Bajorath, 2012). However, the problematic of TyPol is different than these ones because it considers two sets of variables (molecular descriptors and environmental parameters), which are different by nature. Partial least squares regression (PLS) (Wold, 1996; Eriksson et al., 2006) can be used to find the fundamental relation between two sets of variables using a latent variable approach to model the covariance structures in these two spaces. PLS model tries to find the multidimensional directions in the observable variables (i.e. molecular descriptors) space that explain the maximum multidimensional variance direction in the predicted variable (i.e. environmental parameters) space. So PLS, as PCA, constructs uncorrelated variables which summarizes the information, but PLS takes into account the information of both observable and predictive variables. Therefore, the PLS was selected (for a detailed comparison of PLS and PCA, see Maitra and Yan, 2008). After the PLS analysis, a hierarchical clustering algorithm is used to cluster the organic compounds.

2. Materials and methods

2.1. Organic compounds

For this proof-of-concept study of TyPol, 215 organic compounds (191 parent compounds and 24 degradation products) were selected (Tables A1, A2). The selection of these compounds was done according to three criteria: (i) high diversity of chemical families for the parent compounds, (ii) wide ranges of variation of the values of environmental parameters and molecular descriptors (Tables 1, 2), (iii) availability of data for the environmental parameters (see 2.2.). The 191 parent compounds include (i) 116 pesticides taken in the main groups of pesticides (carbamates, organochlorines, organophosphorous, strobilurins, triazines, urea,

phenoxyacids...), (ii) 30 polychlorinated biphenyls (PCB), (iii) 13 polycyclic aromatic hydrocarbons (PAH), (iv) 10 polychlorinated dibenzofurans (PCDF), (v) 9 phthalates, (vi) 7 polychlorinated dibenzodioxins (PCDD), and (vii) 6 miscellaneous compounds (drugs, auxine, hormone…) (Table A1). The ability of TyPol to classify degradation products compared to their parent substance was tested using 24 degradation products deriving from chloride pesticides (Table A2). As some degradation products are common to several parent substances, 26 pairs of parent-degradation product were inputted in TyPol.

2.2. Environmental processes and parameters

Five of the main processes involved in the behavior of organic substances in the environment were retained: (i) dissolution, to describe the expected distribution of the compound between liquid, solid and gaseous phases; (ii) volatilization, which is related to the risk of transfer to atmosphere; (iii) adsorption, which is linked to the risk of transfer to water; (iv) degradation which controls the dissipation and/or the persistence, and increases (or not) the risk of transfer and exposition of a living organism to the substance; and (v) bioaccumulation, to consider the impacts on the organisms and the food chain. Each of these environmental processes can be described by several environmental parameters. In this work, water solubility (*Sw*) and octanol-water partition coefficient (*Kow*) were selected to describe dissolution; vapor pressure (P_{vap}) and Henry's law constant (K_H) for volatilization from soil and plant, and water, respectively; adsorption coefficient normalized to soil carbon organic content (*Koc*) for adsorption; half-life (*DT50*) for degradation; and bioconcentration factor (*BCF*) for ecotoxicity (Table 1). These parameters were chosen because they are the most common ones to represent

the five environmental processes, and because of the availability of the corresponding data in numerous databases.

The values of environmental parameters were mainly taken from the Pesticide Properties DataBase (PPDB, 2013) but also from literature. When values were not available in PPDB (mainly for degradation products), the missing values were collected from Mackay et al. (2006) and ChemSpider (2013). However, considering that a large amount of data of ChemSpider is estimated instead of measured, the use of this database was limited. When several values were available for one environmental parameter, the mean value was retained. For the 215 compounds, 1460 environmental parameters were inputted in TyPol, and there were only 3.9% of missing values. The ranges of values of the parameters are indicated in Table 1 for the 215 compounds.

2.3. Molecular descriptors: selection and calculation

The selection of molecular descriptors was based on a literature review focused on the QSARs that were developed to estimate *Sw*, *Kow*, P_{vap}, K_H, *Koc*, *DT50*, and *BCF*. This review allowed the determination of the molecular descriptors that were best correlated to the seven environmental parameters. We focused on QSARs only built with structural molecular descriptors (number of atoms, molecular surface, dipole moment…) rather than on *Sw* or *Kow*. Indeed, contrary to approaches based on structural molecular descriptors, approaches based on *Sw* or *Kow* are prone to experimental errors in the input variables. However, molecular descriptors accuracy also depends on the approximations chosen to make the calculations. The calibration of the theoretical calculations is driven by the compromise between accuracy and efficiency (Lohninger, 1994; Karelson et al., 1996). Another advantage of the exclusive use of molecular descriptors is that they are calculable for not yet synthesized compounds.

In addition, five criteria were defined to choose the descriptors: (i) their relevance to estimate the seven environmental parameters (see the cited references below), (ii) their common use for the estimation of these seven parameters, (iii) the absence of redundancy between descriptors, (iv) the possibility to calculate the descriptors with molecular modeling, and (v) their ranges of variation. Finally, 40 constitutional, geometrical, topological, and quantum-chemical descriptors were retained (see for example OECD, 1993b; Katritzky et al., 2000; Sabljic, 2001; Dearden and Schüürmann, 2003; Doucette, 2003; Yang et al., 2003; Pavan et al., 2008) (Table 2).

CHEM-3D of ChemOffice Ultra 12.0 (2009) molecular modeling software was used to build three-dimensional chemical structures (3D-structures) in order to calculate the quantum-chemical molecular descriptors (Table 2). As the values of these molecular descriptors are highly dependent on the 3D-structures, a conformational search was done as follows: structures were energy-minimized in MOPAC (Molecular Orbital PACkage) using the semi-empirical method AM1 (Austin Model parameterization) and ground electronic states were obtained as closed-shell molecular orbital wave functions in the restricted Hartree-Fock framework. Analytical frequency calculations have been performed at AM1 level to ensure the obtained structures are minima on the potential energy surface (PES). For each compound, we proceeded by successive steps calculating a large number of conformations deriving from each other by rotations around the different chemical bonds in order to find the global minimum. As first estimate, the descriptors of acido-basic molecules were calculated for their neutral form. The Excel function of ChemOffice was then used to calculate the molecular weights and the Connolly surfaces. Finally, the constitutional (except the molecular weight) and the topological descriptors were calculated

with Dragon 5.5 (2007). For the 215 compounds, 8600 values of molecular descriptors were inputted in TyPol. Their ranges of values are indicated in Table 2.

2.4. Partial least squares (PLS) regression

As stated in the introduction, PLS regression tries to find the multidimensional directions in the observable variables (i.e. structural molecular descriptors) space that explain the maximum multidimensional variance direction in the predicted variable (i.e. environmental parameters) space.

Traditionally, individuals are presented as plots with two components however two axes are not always the optimal choice. Therefore, in this work, the optimal number of axes to perform clustering will be selected using the PRESS (Prediction Sum of Squares) criterion. In addition, PLS can deal with missing values by using the NIPALS (Non-linear Iterative PArtial Least Squares) algorithm. This algorithm allows performing PLS without removing the individuals with missing values and without estimating these missing values (Tenenhaus, 1998). However, the less there are missing values the more accurate the final results are.

2.5. Domain of validity

The knowledge of the domain of validity of the final clustering is important to avoid erroneous conclusions. A priori, TyPol does not have a domain of validity and can be applied to all compounds. However, the use of the PLS algorithm can lead to compounds that are declared atypical by the algorithm. These compounds can be identified using the T^2 of Hotelling (Tenenhaus, 1998). If the T^2 value of a compound is above a calculated threshold, the compound is atypical on the PLS axes. Nevertheless, as one of the objectives of the method is to keep and to

provide the maximum amount of information with the aim of being able to classify a new compound, the clustering of these atypical compounds is done by TyPol.

2.6. Hierarchical clustering

Clustering algorithms are used to assign similar objects into groups (called clusters) based on a similarity criterion chosen by the user. The algorithm used in this study is based on the Ward clustering (Ward, 1963), which keeps the growth of errors as small as possible by merging individuals or clusters. The final number of clusters is chosen after comparison of the heights of the dendrogram, a statistical map which resumes Ward clustering. For the convenience of analyzing clustering of the compounds and their relevant degradation products, arrows linking the parent compounds to their degradation products are represented on the main axes of the PLS. The multivariate analysis is done in R 2.10.0.1 with the "mixOmics" (version 2.8-1) and "cluster" (version 1.13.1) packages. The hierarchical clustering is performed using the agnes() function, and the average linkage and the Euclidean metric are performed under normalized variables (that is mean-centered and scale to unit variance).

2.7. Robustness of the method

To assess the robustness of the clustering method, a classical cross-validation algorithm was used. A fixed percentage of the whole sample is removed from the sample and the PLS is performed. Then, all the 215 compounds (including those which were removed to compute the PLS) are projected on the PLS axes and clustered by the hierarchical clustering algorithm. As compounds which were not included during the PLS algorithm are added in this step, this method can assess the robustness of our methodology and, by consequence, its relevance. Finally, the obtained clustering is compared to the targeted clustering obtained with the PLS

calculated on the whole sample. The closer the clustering is to the targeted one, the more robust the method is because it means that the chemicals which were removed during the PLS step are still well-clustered. We can assume that if the method is robust when substances are removed, it will be relevant when new ones will be tested. This also assesses the predictive quality of the method. The cross-validation study was performed a hundred times using a standard bootstrap procedure for different percentages of removed compounds, and the clusterings were compared using the Adjusted Rand Index (Hubert and Arabie, 1985; Nguyen et al., 2009). This index is a measure of the similarity between two different clusterings. The closer it is to 1 (respectively to 0), the more (respectively less) the two clusterings are similar.

2.8. Computing tools

The information system is based on a management system for relational database MySQL DBMS-R (version 5.1), an Apache web server (version 2.2), and the statistical R software (also used for graphs). The system is installed in a distribution Debian 6.0. The environmental parameters and molecular descriptors are inserted into the management system relational database server which interfaces with Tcl/Tk (Tool command language/Toolkit) made from the R software and "RODBC" library (version 1.3-2). Annotations on the data or results are also stored in the same database. Since the web interfaces are easily editable, statistical analyses of data are treated and helped by the R software Tcl/Tk interfaces. All data that are stored in the DBMS MySQL-R can be viewed via the web interface phpMyAdmin (version 3.3). Data can be imported from phpMyAdmin and new data can easily be inserted. Finally, TyPol was designed in order to easily adapt to other research questions giving the users the choice of the variables (one or several molecular descriptors, one or several environmental parameters), and of the

compounds (all compounds, one or several chemical family…). As the clustering will depend on the needs of the users, no related risk assessment can be included in TyPol.

3. Results and discussion

For this proof-of-concept study, the use of TyPol will be illustrated by clustering the 215 compounds considering all structural molecular descriptors and all environmental parameters. Therefore the results are specific of this case-study (as indicated above, the clustering depends on the needs of the users). The first step in the use of TyPol is the chemical mapping to select the number of components for the subsequent classification, then a hierarchical clustering is performed to identify the optimal number of clusters to classify the organic compounds. For this case-study, some P_{vap}, *Koc*, *DT50* and *BCF* trigger values are proposed to better characterize the clusters (McCall et al., 1980; FOCUS, 2008; Regulation EC 1107/2009, 2009). These trigger values were developed for pesticides, mainly as regulatory threshold values. For the need of our proof-of-concept study, we assumed that they can be extended to any organic compound.

3.1. Chemical mapping by PLS

The choice of the number of PLS components is critical for the subsequent analysis and classification. The number of components which gave the lowest PRESS was therefore selected, it corresponded to the four first axes of the PLS.

The domain of validity of the analysis was studied by calculating the T^2 of Hotelling for the 215 compounds. It appeared that 7 compounds were found as atypical by the four components of the PLS: chlordecone, mirex, kelevan, fosetyl, di-isodecyl, di-isononyl, and benzo(g,h,i)perylene. Indeed, it is well known that these compounds have an extreme behavior in the environment: for example, chlordecone, mirex and kelevan are very persistent (Marchand,

1989; ATSDR, 1995; Cabidoche et al., 2009; Dolfing et al., 2012) contrary to fosetyl which has a very low *DT50* (PPDB, 2013); and di-isodecyl, di-isononyl, and benzo(g,h,i)perylene have very high *Kow* values (PPDB, 2013). Chlordecone, benzo(g,h,i)perylene, mirex and kelevan also have very high connectivity indexes. Nevertheless, these compounds were taken into consideration for the subsequent analysis because they could be representative of other compounds.

The four-component PLS model has good statistical results: R^2_X=0.77, R^2_Y=0.90 and Q^2_Y=0.44. These results shows that the PLS is a good model for the different compounds included in TyPol. The first two components were the most important ones. The closer the compounds are in this score-plot, the more similar they are (Fig. 2). The main characteristic of the first component, which explains 40% of the variance, is the strong positive loadings for all the geometric and topological descriptors, and constitutional descriptors like the number of chlorine or halogen atoms. A contrario, the dipole moment and the total energy have strong negative loadings therefore have an opposite effect. The second axis explains 16% of the variance. On this axis, variables such as the number of chlorine or halogen atoms have a positive loading whereas the number of rotatable, double or simple bonds or the number of hydrogen, oxygen or total atoms have a negative loading (Fig. 3). Figure 3 also shows that many variables seem to be correlated, mainly the different connectivity and valence connectivity indexes. On the third axis, variables such as the number of carbon or hydrogen atoms, and the molecular weight have strong positive loadings, and others such as the number of circuits or the LUMO energy are on the opposite side. On the fourth axis, the HOMO energy and the numbers of rings atoms are on opposite sides of the number of sulfur or chlorine atoms.

3.2. Clustering

Using a hierarchical clustering algorithm, several clusterings, from 1 (all compounds in the same cluster) to 215 (all compounds in a different cluster), were obtained. The selection of the number of clusters is an important and difficult task, which is usually performed by plotting the heights of the dendrogram's node and looking for a break. The results showed that the best choice was to classify the compounds in 6 clusters. The size of the six clusters varied from 3 to 52 compounds (Fig. 2, Table A3), each cluster being characterized by specific features.

Figure 4 shows the range of variations of the values of the 7 environmental parameters for each of the 6 clusters. The importance of the different parameters can be evaluated in Figure 3, but Figure 4 provides a description of the characteristics of each cluster. The trigger value of P_{vap} is indicated to differentiate volatile and non-volatile compounds (there is no trigger value for K_H, FOCUS, 2008), that of *Koc* to differentiate mobile and non-mobile compounds (McCall et al., 1980), that of *DT50* to differentiate persistent and non-persistent compounds (1107/2009 EC, 2009), and that of *BCF* to differentiate compounds having or not a potential of bioaccumulation (1107/2009 EC, 2009) (Fig. 4). Depending on the values of P_{vap}, *Koc*, *DT50* and/or *BCF*, the six clusters aggregate compounds having (or not) risks of air, water and/or soil contamination and/or high ecotoxicity.

The cluster 1 contains 48 compounds and groups together all the thiocarbamates (5 compounds) and nearly 50% of the triazines, carbamates and ureas inputted in TyPol. This cluster is characterized by high values of total energy and polarizability and low values of different connectivity indexes. The compounds have low *Koc* i.e. high risk of groundwater contamination (McCall et al., 1980), and low *DT50* that is low persistence in the environment (Regulation EC 1107/2009, 2009) (Fig. 4). They also have the lowest *BCF* (i.e. low ecotoxicity,

Regulation EC 1107/2009, 2009) and *Kow* among the 6 clusters, which is consistent (Pavan et al., 2008), and the highest *Sw* (this is also consistent with low values of *Kow*) (Fig. 4). Finally, the compounds of cluster 1 have the lowest K_H values (the lowest K_H among the 6 clusters) therefore the lowest volatility from water, but high values of P_{vap} so high volatility from soil and plant, and high risk of transfer to atmosphere (FOCUS, 2008).

Twenty-one of the 30 compounds of cluster 2 are PCB (over 31 inputted in TyPol). There are also 4 organochlorines (2 parent substances and 2 degradation products) and 3 PAH. Compounds of cluster 2 have low dipole moment and high total energy. They also have the lowest *DT50* of the 6 clusters (rapid dissipation), low *Koc* (high mobility), but contrary to cluster 1, low *Sw* (and high *Kow*), and high *BCF* so high ecotoxicity (Fig. 4).

Cluster 3 shares some common traits with cluster 2 in the first two axes. Nevertheless, these two clusters are well separated in the two other axes of the PLS which are not plotted here in a sake of compactness. Cluster 3 is composed of 52 compounds, including all PCDF, 12 organochlorines (9 parent substances and 3 degradation products), 9 PCB, all PCDD, and 10 PAH (13 in the study). The combination of high molecular weights and low number of hydrogen atoms is related to low values of *Sw* (the lowest among the 6 clusters) and high values of *Kow*, and to the highest values of *BCF*. The compounds of cluster 3 also have the highest *DT50* among the six clusters which means very high persistence in soils (Regulation EC 1107/2009, 2009) (Fig. 4). Finally, they have medium values of P_{vap} and K_H (moderate risk of transfer to atmosphere) (Fig. 4).

The cluster 4 contains 37 compounds including all strobilurin compounds, 6 of the 9 phthalates and 5 of the 6 triazoles. The main characteristics of this cluster are very high connectivity indexes, polarizability, and number of hydrogen and carbon atoms for descriptors;

low values of *DT50* and P_{vap}, and medium values of K_H for environmental parameters. The compounds of this cluster have the highest *Koc* values among the 6 clusters, therefore low risk of groundwater contamination (Fig. 4).

Among the 45 compounds of the cluster 5, there are all dinitroanilines, 5 organophosphorous, 4 triazines, 4 urea, and 4 of the 5 chloroacetamides. This cluster is characterized by important dipole moment and number of rotatable bonds for the structural molecular descriptors, and medium values of P_{vap}, K_H, *DT50*, and *Koc*, with high *Sw*, and low *Kow* and *BCF* for the environmental parameters. Few compounds of this cluster are closed to those of cluster 1 in the first two axes of the PLS, but differences between these molecules are more easily noticeable in the fourth axes of the PLS.

Finally, as showed on Figure 2, cluster 6 is an extreme one. It contains mirex, kelevan and chlordecone (in addition, chlordecone is a degradation product of kelevan, PPDB, 2013). As discussed above, these three organochlorine insecticides have very particular chemical structures and high persistence (high *DT50*) in the environment (bishomocubane family). They have extraordinary high values of connectivity or valence connectivity indexes, polarizability, molecular weight, number of chlorine and other halogen atoms; and extremely low values of number of multiple bonds, total energy, HOMO energy. Considering the environmental parameters, they have low *Koc* (high mobility), high *BCF*, that is high ecotoxicity, and high K_H (Dolfing et al., 2012; PPDB, 2013) (Fig. 4). Even on the third and the fourth axes of the PLS, these three compounds have extreme locations and cannot be aggregated with any other cluster.

The other compounds that were detected as atypical by the T^2 of Hotelling are clustered in nearly all the clusters: cluster 1 for fosetyl, cluster 3 for benzo(g,h,i)perylene, and cluster 4 for di-isodecyl and di-isononyl.

The robustness of the method was assessed, using the cross-validation method described above, and found to be high and not depending on a low number of values. The Adjusted Rand Index values were 0.92, 0.87, 0.84 and 0.80 if 1%, 10%, 20% and 50% of the compounds were removed, respectively. As the real cluster of the removed molecules is generally found again, these results show that the predictive quality of the clustering is high. Furthermore, as the molecular descriptors and the environmental parameters were chosen to cover a wide range of values, we can assume that a "new" compound will be clustered with a good quality of prediction. This proof-of-concept study showed that TyPol could allow the classification of organic compounds according to a particular behavior in the environment (i.e. similar values of environmental parameters), which is related to the combination of the values of some specific molecular descriptors.

3.3. Parents-degradation products relationships

To test the ability of TyPol to classify degradation products compared to their parent compounds, 26 pairs of parents and degradation products were inputted (Table A2). The clustering made above using all compounds was retained for the analysis (Table A3). Figure 5 shows the classification of the degradation products compared to their parents. Among all degradation products, 58% (i.e. 15 degradation products) were in the same cluster as their parents. Conversely, 42% (i.e. 11 degradation products) were not in the same cluster as their parents: 6 degradation products originating from parents in clusters 4 and 5 were in cluster 1; 2 degradation products of parent in cluster 3 were in cluster 2; and 3 degradation products of parents in clusters 1 and 4 were in cluster 5. These results are due to similarities (or dissimilarities) in terms of structure and behavior between parent compounds and their

degradation products, but further tests need to be performed with other chemical families. The classification of degradation products compared to the parent compounds will allow the prediction of the behavior in the environment of potential degradation products and/or of degradation products for which no data are available. In addition, the different routes of degradation, i.e. biotic, abiotic (oxidation, dehalogenation…) will be added in the future to investigate if the change in cluster between a compound and its degradation product(s) is related to the type of degradation mechanism.

4. Conclusion

A novel approach, TyPol, for clustering organic compounds according to both their behavior in the environment and their structural molecular descriptors is presented. The approach is based on PLS regression and hierarchical clustering. TyPol considers simultaneously several environmental processes (described by appropriate environmental parameters), and the degradation products of compounds.

This proof-of-concept study, based on the classification of 215 organic compounds, showed that the combination of the values of some molecular descriptors could be related to a particular behavior in the environment. The robustness of the method was studied and demonstrated to be good, as well as the statistical performances of the PLS regression. Therefore, TyPol could help to predict the environmental behavior of a "new" compound from its affiliation to one cluster or to select representative substances from a large data set in order to answer some specific questions regarding their behavior in the environment. In addition, TyPol takes into account the degradation products of organic compounds. The analysis is based on the same methodology as above and highlights the similarities (or dissimilarities) between a parent substance and its degradation product. One of the next steps of this work will investigate if the

change in cluster between a compound and its degradation product(s) is related to the type of degradation mechanism (oxidation, epoxidation, hydroxylation…). Additional environmental and ecotoxicological parameters, and molecular descriptors will also be included in TyPol to refine the classification of compounds.

Acknowledgements

The authors acknowledge the Projet Innovant of the "Environnement et Agronomie" Department of INRA and the AIP DEMICHLORD (Etudes exploratoires de la dégradation microbienne de la chlordécone) of INRA for financial supports, and the FIRE (Fédération Ile-de-France de Recherche sur l'Environnement) for Ziang Li's grant. They are also grateful to Anaïs Labrunie and Sophie Vitrant for their contribution to this work. Finally, the authors would like to thank the anonymous reviewers for their constructive comments.

Appendix A. Supplementary material

Supplementary Tables A1, A2, A3

References

Ahlers, J., Stock, F., Werschkun, B., 2008. Integrated testing and intelligent assessment-new challenges under REACH. Environ. Sci. Pollut. Res. 15, 565-572.

ATSDR (Agency for Toxic Substances and Disease Registry), 1995. Toxicological profile for mirex and chlordecone. US Department of Health and Human Services, Public Health Services, http://www.atsdr.cdc.gov/toxprofiles/tp66.pdf

Cabidoche, Y.-M., Achard, R., Cattan, P., Clermont-Dauphin, C., Massat, F., Sansoulet, J., 2009. Long-term pollution by chlordecone of tropical volcanic soils in the French West Indies: A simple leaching model accounts for current residue. Environ. Pollut. 157, 1697-1705.

Carsel, R.F., Imhoff, J.C., Hummel, P.R., Cheplick, J.M., Donigian Jr, A.S., 1998. PRZM-3: a Model for Predicting Pesticide and Nitrogen Fate in the Crop Root and Unsaturated Soil Zones: Users Manual for Release 3.12 National Exposure Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Athens, GA.

ChemOffice, 2009. ChemOffice Ultra 12.0 molecular modelling software, Cambridge Soft, Perkin Elmer.

ChemSpider, 2013. The free chemical database. http://www.chemspider.com/

Dearden, J.C., Schüürmann, G., 2003. Quantitative structure-property relationships for predicting Henry's law constant from molecular structure. Environ. Toxicol. Chem. 22, 1755-1770.

Dolfing, J., Novak, I., Archelas, A., Macarie, H., 2012. Gibbs free energy of formation of chlordecone and potential degradation products: implications for remediation strategies and environmental fate. Environ. Sci. Technol. 46, 8131-8139.

Doucette, W.J., 2003. Quantitative structure-activity relationships for predicting soil-sediment sorption coefficients for organic chemicals. Environ. Toxicol. Chem. 22, 1771-1788.

Dragon 5.5, 2007. Software for the calculation of molecular descriptors, Talete s.r.l. http://www.talete.mi.it/

Eriksson, L., Andersson, P.L., Johansson, E., Tysklind, M., 2002. Multivariate biological profiling and principal toxicity regions of compounds: the PCB case study. J. Chemometrics 16, 497-509.

Eriksson, L., Andersson, P., Johansson, E., Tysklind, M., 2006. Megavariate analysis of environmental QSAR data. Part I - A basic framework founded on principal component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD). Mol. Divers. 10, 169-186.

FOCUS, 2008. Pesticides in air: considerations for exposure assessment. Report of the FOCUS Working Group on Pesticides in Air, EC Document Reference SANCO/10553/2006 Rev 2, June 2008.

Gramatica, P., Papa, E., Battaini,F., 2004. Ranking and classification of non-ionic organic pesticides for environmental distribution: a QSAR approach. Int. J. Env. Anal. Chem. 84, 65-74.

Harju, M., Andersson, P.L., Haglund, P., Tysklind, M., 2002. Multivariate physicochemical characterisation and quantitative structure-property relationship modeling of polybrominated diphenyl ethers. Chemosphere 47, 375-384.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193-218.

Jackson, J.E.A., 1991. User's Guide to Principal Components, eds. John Wiley and Sons, New York.

Jarvis, N., Larsbo, M., 2012. MACRO (v5.2): model use, calibration and validation. Trans. ASABE 55, 1413-1423.

Karelson, M., Lobanov, V.S., Katritzky, A.R., 1996. Quantum-chemical descriptors in QSAR/QSPR studies. Chem. Rev. 96, 1027-1043.

Katritzky, A.R., Maran, U., Lobanov, V.S., Karelson, M., 2000. Structurally diverse quantitative structure-property relationship correlations of technologically relevant physical properties. J. Chem. Inf. Comput. Sci. 40, 1-18

1 Lohninger, H., 1994. Estimation of soil partition coefficients of pesticides from their chemical
2 structure. Chemosphere 29, 1611-1626.
3 Mackay, D., McCarty, L.S., McLeod, M., 2001. On the validity of classifying chemicals for
4 persistence, bioaccumulation, toxicity, and potential for long-range transport. Environ.
5 Tox. Chem. 20, 1491-1498.
6 Mackay, D., Shiu, W.Y., Ma, K.-C., Lee, S.C., 2006. Handbook of physical-chemical properties
7 and environmental fate for organic chemicals, second ed. CRC Press, Taylor and Francis
8 Group, Boca Raton.
9 McCall, P.J., Swann, R.L., Laskowski, D.A., Unger, S.M., Vrona, S.A., Dishburger, H.J., 1980.
10 Estimation of chemical mobility in soil from liquid chromatographic retention times. Bull.
11 Environ. Contam. Toxicol. 24, 190-195.
12 Maitra, S., Yan, J., 2008. Principal component analysis and partial least squares: two-dimension
13 reduction techniques for regression. Proceedings of the Casual Actuarial Society, 79-90.
14 Marchand, A.P., 1989. Synthesis and chemistry of homocubanes, bishomocubanes and
15 trishomocubanes. Chem. Rev. 89, 1011-1033.
16 Mishra, M., Potetz, B., Huan, J., 2011. Bayesian classifiers for chemical toxicity prediction.
17 BIBM'11, Proceedings of the 2011 IEEE International Conference on Bioinformatics and
18 Biomedicine, 595-599.
19 Muir, D.C.G., Howard, P.H., 2006. Are there other persistent organic pollutants ? A challenge
20 for environmental chemists. Environ. Sci. Technol. 40, 7157-7166.
21 Nguyen, X., Epps, J., Bailey, J., 2009. Information theoretic measures for clustering comparison:
22 Is a correction for chance necessary? ICML'09: Proceedings of the 26th Annual
23 International Conference on Machine Learning, San Francisco, 1073-1080.

OECD (Organisation for economic co-operation and development), 1993a. Application of structure-activity relationships to the estimation of properties important in exposure assessment. Environment monographs No 67, OECD, Paris.

OECD (Organisation for economic co-operation and development), 1993b. Structure-activity relationships for biodegradation. Environment monograph No 68, OECD, Paris.

Pavan, M., Netzeva, T.I., Worth, A.P., 2008. Review of literature-based quantitative structure-activity relationship models for bioconcentration. QSAR Comb. Sci. 27, 21-31.

PPDB (Pesticide properties database), 2013. http://sitem.herts.ac.uk/aeru/footprint/index2.htm

Rännar, S., Andersson, P.L., 2010. A novel approach using hierarchical clustering to select industrial chemicals for environmental impact assessment. J. Chem. Inf. Mod. 50, 30-36.

Rännar, S., Andersson, P.L., 2011. A multivariate chemical similarity approach to search for drugs of potential environmental concern. J. Chem. Inf. Mod. 51, 1788-1794.

Raymond, J.W., Rogers, T.N., Shonnard, D.R., Kline, A.A., 2001. A review of structure-based biodegradation estimation methods. J. Hazard. Mater. B84, 189-215.

Regulation EC 1107/2009, 2009. Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC.

Sabljic, A., 2001. QSAR models for estimating properties of persistent organic pollutants required in evaluation of their environmental fate and risk. Chemosphere 43, 363-375.

Saito, S., Ohno, K., Sese, J., Sugarawa, K., Sakuraba, H., 2010. Prediction of the clinical phenotype of Fabry disease based on protein sequential and structural information. J. Hum. Genet. 55, 175-178.

Snarey, M., Terrett, N.K., Willet, P., Wilton, D.J., 1997. Comparison of algorithms for dissimilarity-based compound selection. J. Mol. Graph. Model. 15, 372-385.

Tenenhaus, M., 1998. La regression PLS, Théorie et Pratique, ed. Technip, Paris.

Vogt, M., Bajorath J., 2012. Chemoinformatics: a view of the field and current trends in method development. Bio. Med. Chem. 20, 5317-5323.

Walker, J.D., Carlsen, L., 2002. QSARs for identifying and prioritizing substances with persistence and bioconcentration potential. SAR QSAR Environ. Res. 13, 713-725.

Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236-244.

Wold, H., 1996. Estimation of principal component and related models by iterative least squares, in: Krishnaiaah, P.R. (Ed.), Multivariate Analysis, Academic Press, New York, pp. 391-420.

Worrall, F., 2001. A molecular topology approach to predicting pesticide pollution of groundwater. Environ. Sci. Technol. 35, 2282-2287.

Yang, P., Chen. J., Chen, S., Yuan, X., Schramm, K.-W., Kettrup, A., 2003. QSPR models for physicochemical properties of polychlorinated diphenyl ethers. Sci. Tot. Environ. 305, 65-76.

Figure captions

Fig 1. Construction and structure of TyPol. A "new" compound is either a parent compound or a degradation product, for which the environmental parameters (for example water solubility *Sw*, adsorption coefficient *Koc* or bioconcentration factor *BCF*) are not known.

Fig. 2. Clustering of the 215 organic compounds in six clusters (each cluster has a different symbol) on the two main components of the PLS (PLS1 and PLS2).

Fig. 3. Circles of correlations of the "environmental parameters" (in blue) and "molecular descriptors" (in red) variables on the two main components of the PLS (PLS1 and PLS2). C.i-i stands for the connectivity index C.i of order i (i = 0 to 5), and V.c-i stands for the valence connectivity index V.c of order i (i = 0 to 5).

Fig. 4. Range of variation (box-and-whisker plots) of the values of the seven environmental parameters considered into TyPol (water solubility *Sw*, octanol-water partition coefficient *Kow*, vapor pressure P_{vap}, Henry's law constant K_H, adsorption coefficient normalized to soil carbon organic content *Koc*, half-life *DT50*, and bioconcentration factor *BCF*) for each cluster after analysis of the 215 organic compounds. Dotted lines represent the limits between: volatile (log $P_{vap} > -1$) and non-volatile compounds (log $P_{vap} < -1$) (FOCUS, 2008); mobile (log *Koc* < 2.7) and non-mobile compounds (log *Koc* > 2.7) (McCall et al., 1980); persistent (log *DT50* > 2.25) and non-persistent compounds (log *DT50* < 2.25) (Regulation EC 1107/2009, 2009), and compounds having (log *BCF* > 2) or not (log *BCF* < 2) a potential of bioaccumulation (Regulation EC 1107/2009, 2009).

Fig. 5. Relationships between the parent compounds and their degradation products on the two main components of the PLS (PLS1 and PLS2). Arrows are drawn from the parent compounds (the CAS number is indicated) to their degradation products.

Fig. 1.

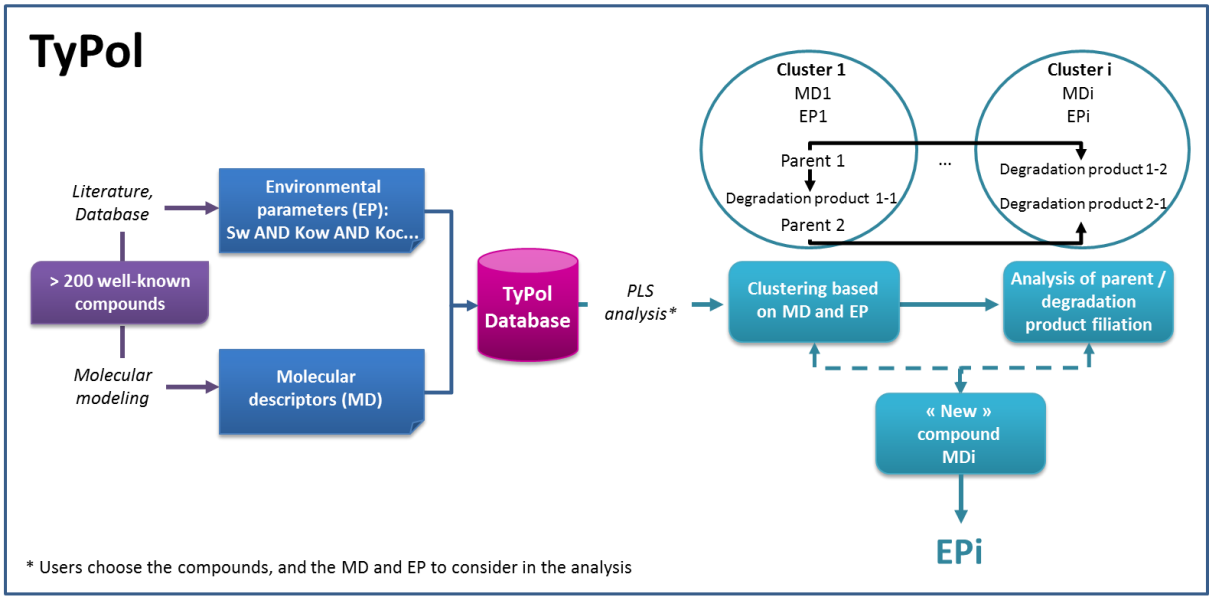

TyPol
Literature, Database
Environmental parameters (EP): Sw AND Kow AND Koc...
> 200 well-known compounds
Molecular modeling
Molecular descriptors (MD)
TyPol Database
PLS analysis*
Clustering based on MD and EP
Cluster 1
MD1
EP1
Parent 1
Degradation product 1-1
Parent 2
Cluster i
MDi
EPi
Degradation product 1-2
Degradation product 2-1
Analysis of parent / degradation product filiation
« New » compound MDi
EPi
* Users choose the compounds, and the MD and EP to consider in the analysis

Fig. 2.

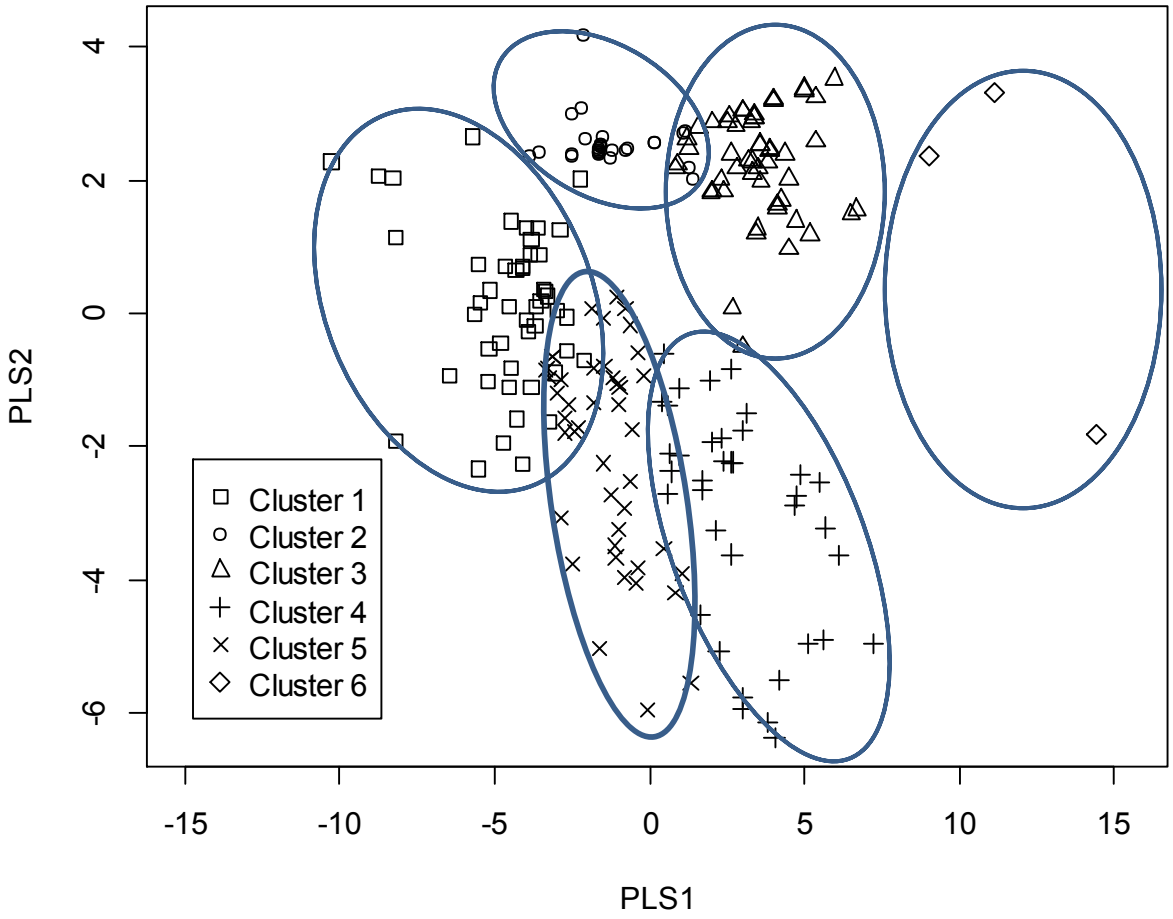
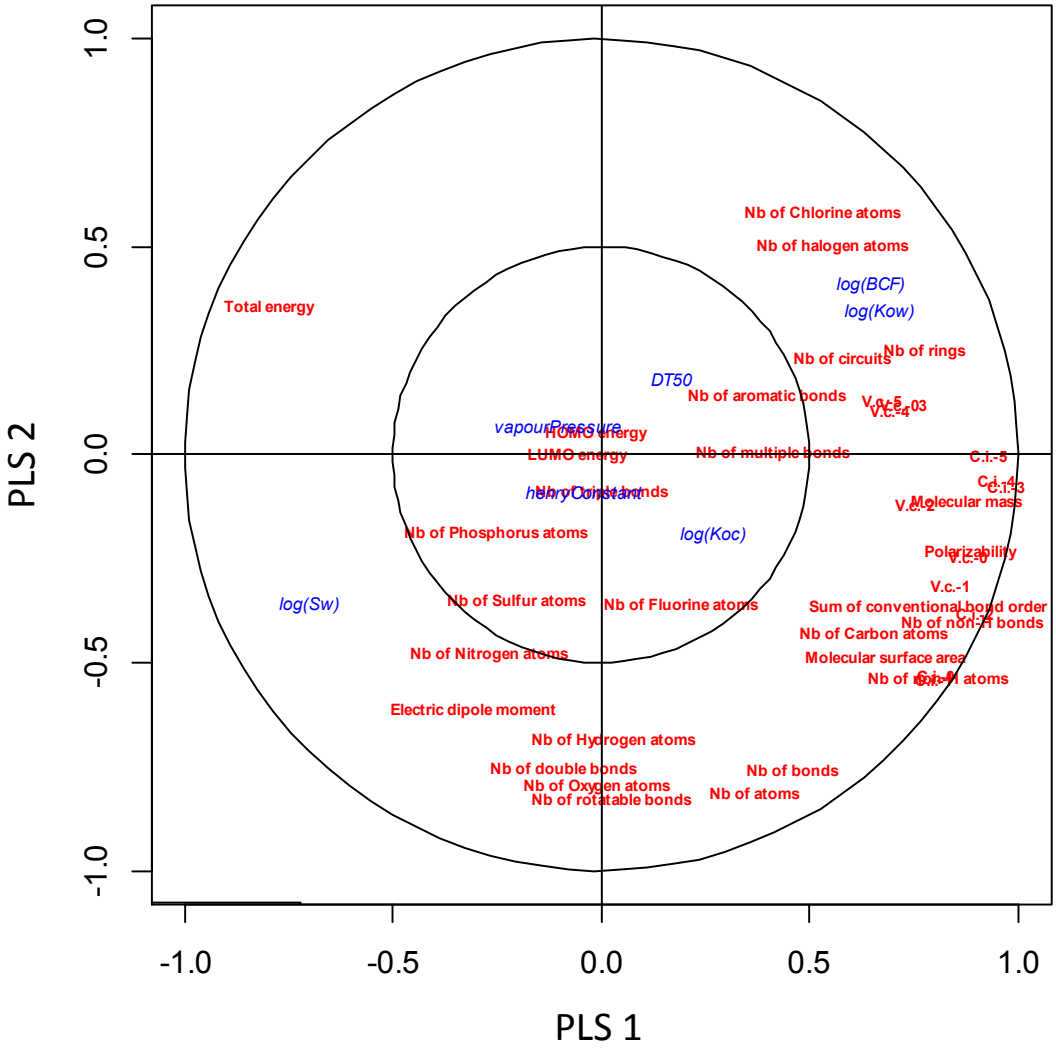
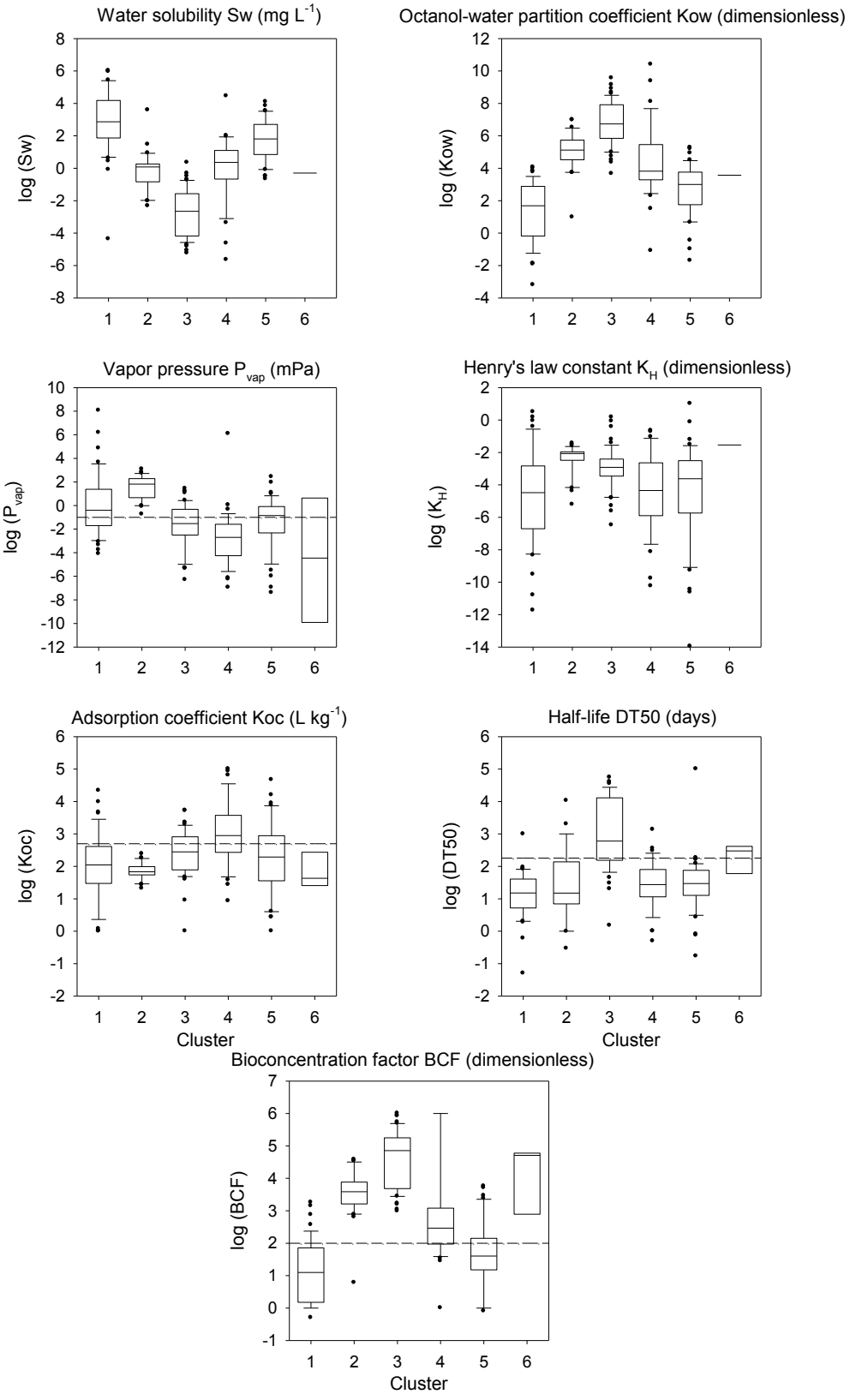

PLS2
PLS1
4
2
0
-2
-4
-6
-15
-10
-5
0
5
10
15
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

Fig. 3.

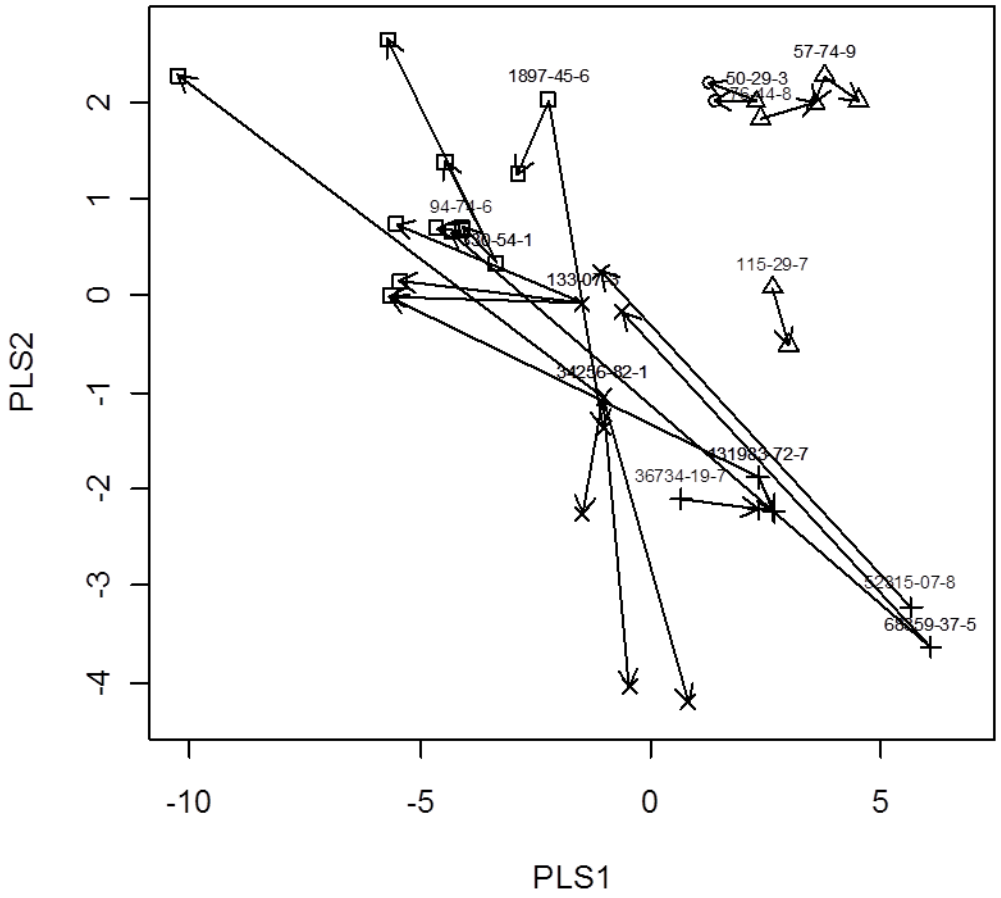

PLS 2
PLS 1
1.0
0.5
0.0
-0.5
-1.0
Nb of Chlorine atoms
Nb of halogen atoms
log(BCF)
log(Kow)
Total energy
Nb of circuits
Nb of rings
DT50
Nb of aromatic bonds
V.c.-5
V.c.-4
V.c.-3
vapourPressure
HOMO energy
LUMO energy
Nb of multiple bonds
C.i.-5
C.i.-4
C.i.-3
henryConstant
Molecular mass
V.c.-2
Nb of Phosphorus atoms
log(Koc)
Polarizability
V.c.-0
V.c.-1
log(Sw)
Nb of Sulfur atoms
Nb of Fluorine atoms
Sum of conventional bond order
Nb of non-H bonds
Nb of Carbon atoms
Nb of Nitrogen atoms
Molecular surface area
Nb of non-H atoms
Electric dipole moment
Nb of Hydrogen atoms
Nb of double bonds
Nb of bonds
Nb of Oxygen atoms
Nb of atoms
Nb of rotatable bonds

1 **Fig. 4.**


Water solubility Sw (mg L-1)
log (Sw)
Octanol-water partition coefficient Kow (dimensionless)
log (Kow)
Vapor pressure Pvap (mPa)
log (Pvap)
Henry's law constant KH (dimensionless)
log (KH)
Adsorption coefficient Koc (L kg-1)
log (Koc)
Half-life DT50 (days)
log (DT50)
Cluster
Cluster
Bioconcentration factor BCF (dimensionless)
log (BCF)
Cluster

2

Fig. 5.


PLS2
PLS1
-10
-5
0
5
2
1
0
-1
-2
-3
-4
57-74-9
1897-45-6
50-29-3
94-74-6
115-29-7
34256-82-1
431983-72-7
36734-19-7
52315-07-8
68359-37-5

Tables

Table 1

Environmental parameters and ranges of variation of their values for the 215 organic compounds (191 parent compounds and 24 degradation products) inputted in TyPol (*Sw*: water solubility, *Kow*: octanol-water partition coefficient, P_{vap}: vapor pressure, K_H: Henry's law constant, *Koc*: adsorption coefficient, *DT50*: half-life, *BCF*: bioconcentration factor)

Environmental process	Environmental parameter	Range of variation	
		Parent compound	Degradation product
Dissolution	log [*Sw* (mg L^{-1})]	[- 6.72 ; 10.29]	[- 3.63 ; 13.82]
	log [*Kow* (dimensionless)]	[- 0.81 ; 6.91]	[- 0.17 ; 6.51]
Volatilization	P_{vap} (mPa)	[5×10^{-5} ; 27]	[5×10^{-4} ; 1×10^{8}]
	K_H (dimensionless)	[2×10^{-6} ; 1.48]	[1.1×10^{-14} ; 1.48]
Adsorption	log [*Koc* (L kg^{-1})]	[2.19 ; 11.36]	[1 ; 6.83]
Degradation	*DT50* (days)	[4.7 ; 4100]	[0.05 ; 10603]
Ecotoxicity	log [*BCF* (dimensionless)]	[0 ; 4]	[0 ; 3.93]

Table 2

List of the 40 molecular descriptors inputted in TyPol and ranges of variation of their values for the 215 organic compounds (191 parent compounds and 24 degradation products)

Category	Molecular descriptor	Range of variation	
		Parent compound	Degradation product
Constitutional	Number of atoms	[14 ; 47]	[8 ; 46]
	Number of non-H atoms	[13 ; 29]	[4 ; 23]
	Number of hydrogen atoms	[0 ; 20]	[1 ; 23]
	Number of carbon atoms	[8 ; 22]	[2 ; 17]
	Number of nitrogen atoms	[0 ; 3]	[0 ; 3]
	Number of oxygen atoms	[0 ; 3]	[0 ; 5]
	Number of phosphorus atoms	[0 ; 0]	[0 ; 0]
	Number sulfur atoms	[0 ; 1]	[0 ; 1]
	Number fluorine atoms	[0 ; 1]	[0 ; 1]
	Number of chlorine atoms	[1 ; 8]	[0 ; 8]
	Number of halogen atoms	[1 ; 8]	[0 ; 8]
	Number of bonds	[14 ; 49]	[7 ; 46]
	Number of non-H bonds	[13 ; 31]	[3 ; 25]
	Number of double bonds	[0 ; 3]	[0 ; 3]
	Number of triple bonds	[0 ; 2]	[0 ; 2]
	Number of multiple bonds	[1 ; 15]	[1 ; 13]
	Number of rotatable bonds	[0 ; 7]	[0 ; 9]
	Number of aromatic bonds	[0 ; 12]	[0 ; 12]
	Sum of conventional bond order	[17 ; 41]	[4 ; 31.5]
	Number of rings	[1 ; 3]	[0 ; 4]
	Number of circuits	[1 ; 6]	[0 ; 10]
	Molecular weight (g mol^{-1})	[200 ; 434.3]	[60 ; 423.76]
Geometric	Connolly molecular surface area ($Å^2$)	[186.1 ; 311.6]	[73.46 ; 278.7]
Topological	Connectivity index of order 0	[9.84 ; 21.18]	[3.57 ; 17.41]
	Connectivity index of order 1	[6.09 ; 13.73]	[1.73 ; 10.92]
	Connectivity index of order 2	[5.58 ; 13.02]	[1.73 ; 10.70]
	Connectivity index of order 3	[3.72 ; 9.94]	[0 ; 10.26]
	Connectivity index of order 4	[2.67 ; 7.96]	[0 ; 8.10]
	Connectivity index of order 5	[2 ; 6.40]	[0 ; 6.78]
	Valence connectivity index of order 0	[7.83 ; 17.25]	[2.36 ; 14.79]
	Valence connectivity index of order 1	[4.08 ; 9.65]	[0.93 ; 8.96]
	Valence connectivity index of order 2	[3.03 ; 9.29]	[0.52 ; 9.68]
	Valence connectivity index of order 3	[1.72 ; 9.79]	[0 ; 10.13]
	Valence connectivity index of order 4	[1.11 ; 7.46]	[0 ; 7.67]
	Valence connectivity index of order 5	[0.63 ; 5.95]	[0 ; 6.23]
Quantum-chemical	Polarizability ($Å^3$)	[19.89 ; 45.58]	[5.13 ; 45.58]
	Electric dipole moment (D)	[1.07 ; 5.46]	[0.07 ; 4.83]
	HOMO energy (eV)	[- 10.35 ; - 8.95]	[- 11.62 ; - 8.73]
	LUMO energy (eV)	[- 1.76 ; 0.05]	[- 2.20 ; 0.98]
	Total energy (eV)	[- 5462 ; - 2611]	[- 5462 ; - 953]

Supplementary material

Table A1

List of organic compounds inputted in TyPol (chemical families, CAS numbers, names and chemical formulae) (PCB: polychlorinated biphenyls, PAH: polycyclic aromatic hydrocarbons, PCDF: polychlorinated dibenzofurans, PCDD: polychlorinated dibenzodioxins)

Chemical family	CAS number	Name	Chemical formula	Chemical family	CAS number	Name	Chemical formula
Pesticides							
Organochlorine	**50-29-3**	p,p'-DDT	$C_{14}H_9Cl_5$	PCB	**92-52-4**	Biphenyl	$C_{12}H_{10}$
Organochlorine	**57-74-9**	Chlordane	$C_{10}H_6Cl_8$	PCB	**2050-67-1**	3,3'-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**58-89-9**	Lindane	$C_6H_6Cl_6$	PCB	**2050-68-2**	4,4'-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**60-57-1**	Dieldrine	$C_{12}H_8Cl_6O$	PCB	**2051-24-3**	Decachlorobiphenyl	$C_{12}Cl_{10}$
Organochlorine	**72-20-8**	Endrine	$C_{12}H_8Cl_6O$	PCB	**2051-60-7**	2-chlorobiphenyl	$C_{12}H_9Cl$
Organochlorine	**76-44-8**	Heptachlore	$C_{10}H_5Cl_7$	PCB	**2051-61-8**	3-chlorobiphenyl	$C_{12}H_9Cl$
Organochlorine	**115-29-7**	Endosulfan	$C_9H_6Cl_6O_3S$	PCB	**2974-90-5**	3,4'-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**118-74-1**	Hexachlorobenzene	C_6Cl_6	PCB	**2974-92-7**	3,4-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**143-50-0**	Chlordecone	$C_{10}Cl_{10}O$	PCB	**13029-08-8**	2,2'-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**297-78-9**	Isobenzane	$C_9H_4Cl_8O$	PCB	**16605-91-7**	2,3-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**1715-40-8**	Bromocyclene	$C_8H_5BrCl_6$	PCB	**25569-80-6**	2,3'-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**2385-85-5**	Mirex	$C_{10}Cl_{12}$	PCB	**33146-45-1**	2,6-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**4234-79-1**	Kelevan	$C_{17}H_{12}Cl_{10}O_4$	PCB	**33284-50-3**	2,4-dichlorobiphenyl	$C_{12}H_8Cl_2$
Organochlorine	**8001-35-2**	Toxaphene	$C_{10}H_8Cl_8$	PCB	**34883-39-1**	2,5-dichlorobiphenyl	$C_{12}H_8Cl_2$
Carbamate	**63-25-2**	Carbaryl	$C_{12}H_{11}NO_2$	PCB	**34883-41-5**	3,5-dichlorobiphenyl	$C_{12}H_8Cl_2$
Carbamate	**101-21-3**	Chlorpropham	$C_{10}H_{12}ClNO_2$	PCB	**34883-43-7**	2,4'-dichlorobiphenyl	$C_{12}H_8Cl_2$
Carbamate	**116-06-3**	Aldicarb	$C_7H_{14}N_2O_2S$	PCB	**35065-30-6**	2,2',3,3',4,4',5-heptachlorobiphenyl	$C_{12}H_3Cl_7$
Carbamate	**1563-66-2**	Carbofuran	$C_{12}H_{15}NO_3$	PCB	**35694-08-7**	2,2',3,3',4,4',5,5'-octachlorobiphenyl	$C_{12}H_2Cl_8$
Carbamate	**3337-71-1**	Asulam	$C_8H_{10}N_2O_4S$	PCB	**37680-66-3**	2,2',4-trichlorobiphenyl	$C_{12}H_7Cl_3$
Carbamate	**13684-56-5**	Desmedipham	$C_{16}H_{16}N_2O_4$	PCB	**38380-07-3**	2,2',3,3',4,4'-hexachlorobiphenyl	$C_{12}H_4Cl_6$
Carbamate	**13684-63-4**	Phenmedipham	$C_{16}H_{16}N_2O_4$	PCB	**38444-78-9**	2,2',3-trichlorobiphenyl	$C_{12}H_7Cl_3$
Carbamate	**16118-49-3**	Carbetamide	$C_{12}H_{16}N_2O_3$	PCB	**38444-93-8**	2,2',3,3'-tetrachlorobiphenyl	$C_{12}H_6Cl_4$
Carbamate	**23103-98-2**	Pirimicarb	$C_{11}H_{18}N_4O_2$	PCB	**40186-72-9**	2,2',3,3',4,4',5,5',6-nonachlorobiphenyl	$C_{12}HCl_9$
Carbamate	**23135-22-0**	Oxamyl	$C_7H_{13}N_3O_3S$	PCB	**52663-59-9**	2,2',3,4-tetrachlorobiphenyl	$C_{12}H_6Cl_4$
Carbamate	**24579-73-5**	Propamocarb	$C_9H_{20}N_2O_2$	PCB	**52663-62-4**	2,2',3,3',4-pentachlorobiphenyl	$C_{12}H_5Cl_5$
Carbamate	**79127-80-3**	Fenoxycarb	$C_{17}H_{19}NO_4$	PCB	**52663-71-5**	2,2',3,3',4,4',6-heptachlorobiphenyl	$C_{12}H_3Cl_7$

Urea	**101-42-8**	Fenuron	$C_9H_{12}N_2O$	PCB	**52663-78-2**	2,2',3,3',4,4',5,6-octachlorobiphenyl	$C_{12}H_2Cl_8$
Urea	**330-54-1**	Diuron	$C_9H_{10}Cl_2N_2O$	PCB	**52663-79-3**	2,2',3,3',4,4',5,6,6'-nonachlorobiphenyl	$C_{12}HCl_9$
Urea	**330-55-2**	Linuron	$C_9H_{10}Cl_2N_2O_2$	PCB	**55215-18-4**	2,2',3,3',4,5-hexachlorobiphenyl	$C_{12}H_4Cl_6$
Urea	**555-37-3**	Neburon	$C_{12}H_{16}Cl_2N_2O$	PCB	**60145-20-2**	2,2',3,3',5-pentachlorobiphenyl	$C_{12}H_5Cl_5$
Urea	**1746-81-2**	Monolinuron	$C_9H_{11}ClN_2O_2$				
Urea	**3060-89-7**	Metobromuron	$C_9H_{11}BrN_2O_2$	PAH	**50-32-8**	Benzo(a)pyrene	$C_{20}H_{12}$
Urea	**15545-48-9**	Chlortoluron	$C_{10}H_{13}ClN_2O$	PAH	**53-70-3**	Dibenzo(a,h)anthracene	$C_{22}H_{14}$
Urea	**34123-59-6**	Isoproturon	$C_{12}H_{18}N_2O$	PAH	**56-55-3**	Benzo(a)anthracene	$C_{18}H_{12}$
Urea	**64902-72-3**	Chlorsulfuron	$C_{12}H_{12}ClN_5O_4S$	PAH	**85-01-8**	Phenanthrene	$C_{14}H_{10}$
Urea	**79510-48-8**	Metsulfuron	$C_{13}H_{13}N_5O_6S$	PAH	**86-73-7**	Fluorene	$C_{13}H_{10}$
Organophosphorous	**56-38-2**	Parathion-ethyl	$C_{10}H_{14}NO_5PS$	PAH	**120-12-7**	Anthracene	$C_{14}H_{10}$
Organophosphorous	**60-51-5**	Dimethoate	$C_5H_{12}NO_3PS_2$	PAH	**129-00-0**	Pyrene	$C_{16}H_{10}$
Organophosphorous	**86-50-0**	Azinphos-methyl	$C_{10}H_{12}N_3O_3PS_2$	PAH	**191-24-2**	Benzo(g,h,i)perylene	$C_{22}H_{12}$
Organophosphorous	**121-75-5**	Malathion	$C_{10}H_{19}O_6PS_2$	PAH	**193-39-5**	Indeno[1,2,3-cd]pyrene	$C_{22}H_{12}$
Organophosphorous	**122-14-5**	Fenitrothion	$C_9H_{12}NO_5PS$	PAH	**205-99-2**	Benzo(b)fluoranthene	$C_{20}H_{12}$
Organophosphorous	**333-41-5**	Diazinon	$C_{12}H_{21}N_2O_3PS$	PAH	**206-44-0**	Fluoranthene	$C_{16}H_{10}$
Organophosphorous	**5598-13-0**	Chlorpyrifos-methyl	$C_7H_7Cl_3NO_3PS$	PAH	**207-08-9**	Benzo(k)fluoranthene	$C_{20}H_{12}$
Organophosphorous	**13194-48-4**	Ethoprophos	$C_8H_{19}O_2PS_2$	PAH	**218-01-9**	Chrysene	$C_{18}H_{12}$
Organophosphorous	**15845-66-6**	Fosetyl	$C_2H_7O_3P$				
Amide	**709-98-8**	Propanil	$C_9H_9Cl_2NO$	PCDF	**39001-02-0**	OCDF	$C_{12}Cl_8O$
Amide	**15299-99-7**	Napropamide	$C_{17}H_{21}NO_2$	PCDF	**51207-31-9**	2,3,7,8-TCDF	$C_{12}H_4Cl_4O$
Amide	**23950-58-5**	Propyzamide	$C_{12}H_{11}Cl_2NO$	PCDF	**55673-89-7**	1,2,3,4,7,8,9-HpCDF	$C_{12}HCl_7O$
Amide	**35256-85-0**	Tebutam	$C_{15}H_{23}NO$	PCDF	**57117-31-4**	2,3,4,7,8-PeCDF	$C_{12}H_3Cl_5O$
Amide	**55814-41-0**	Mepronil	$C_{17}H_{19}NO_2$	PCDF	**57117-41-6**	1,2,3,7,8-PeCDF	$C_{12}H_3Cl_5O$
Amide	**57837-19-1**	Metalaxyl	$C_{15}H_{21}NO_4$	PCDF	**57117-44-9**	1,2,3,6,7,8-HxCDF	$C_{12}H_2Cl_6O$
Amide	**77732-09-3**	Oxadixyl	$C_{14}H_{18}N_2O_4$	PCDF	**60851-34-5**	2,3,4,6,7,8-HxCDF	$C_{12}H_2Cl_6O$
Amide	**180409-60-3**	Cyflufenamid	$C_{20}H_{17}F_5N_2O_2$	PCDF	**67562-39-4**	1,2,3,4,6,7,8-HpCDF	$C_{12}HCl_7O$
Strobilurin	**117428-22-5**	Picoxystrobin	$C_{18}H_{16}F_3NO_4$	PCDF	**70648-26-9**	1,2,3,4,7,8-HxCDF	$C_{12}H_2Cl_6O$
Strobilurin	**131860-33-8**	Azoxystrobin	$C_{22}H_{17}N_3O_5$	PCDF	**72918-21-9**	1,2,3,7,8,9-HxCDF	$C_{12}H_2Cl_6O$
Strobilurin	**141517-21-7**	Trifloxystrobin	$C_{20}H_{19}F_3N_2O_4$				
Strobilurin	**143390-89-0**	Kresoxim-methyl	$C_{18}H_{19}NO_4$	Phthalate	**84-61-7**	Di-cyclohexyl	$C_{20}H_{26}O_4$
Strobilurin	**149961-52-4**	Dimoxystrobin	$C_{19}H_{22}N_2O_3$	Phthalate	**84-66-2**	Diethyl	$C_{12}H_{14}O_4$
Strobilurin	**175013-18-0**	Pyraclostrobin	$C_{19}H_{18}ClN_3O_4$	Phthalate	**84-74-2**	Dibutyl	$C_{16}H_{22}O_4$
Strobilurin	**361377-29-9**	Fluoxastrobin	$C_{21}H_{16}ClFN_4O_5$	Phthalate	**85-68-7**	Benzylbutyl	$C_{19}H_{20}O_4$
Triazine	**122-34-9**	Simazine	$C_7H_{12}ClN_5$	Phthalate	**117-81-7**	Di-2-ethylhexyl	$C_{24}H_{38}O_4$
Triazine	**834-12-8**	Ametryn	$C_9H_{17}N_5S$	Phthalate	**117-84-0**	Di-n-octyl	$C_{24}H_{38}O_4$
Triazine	**886-50-0**	Terbutryn	$C_{10}H_{19}N_5S$	Phthalate	**131-11-3**	Dimethyl	$C_{10}H_{10}O_4$
Triazine	**1912-24-9**	Atrazine	$C_8H_{14}ClN_5$	Phthalate	**26761-40-0**	Di-isodecyl	$C_{28}H_{46}O_4$
Triazine	**5915-41-3**	Terbuthylazine	$C_9H_{16}ClN_5$	Phthalate	**28553-12-0**	Di-isononyl	$C_{26}H_{42}O_4$
Triazine	**21725-46-2**	Cyanazine	$C_9H_{13}ClN_6$				
Triazine	**66215-27-8**	Cyromazine	$C_6H_{10}N_6$	PCDD	**1746-01-6**	2,3,7,8-tetrachloro-dibenzo-p-dioxine	$C_{12}H_4Cl_4O_2$
Diazine	**1698-60-8**	Chloridazon	$C_{10}H_8ClN_3O$	PCDD	**3268-87-9**	OCDD	$C_{12}Cl_8O_2$

Diazine	**25057-89-0**	Bentazone	$C_{10}H_{12}N_2O_3S$	PCDD	**19408-74-3**	1,2,3,7,8,9-HxCDD	$C_{12}H_2Cl_6O_2$
Triazinone	**21087-64-9**	Metribuzin	$C_8H_{14}N_4OS$	PCDD	**35822-46-9**	1,2,3,4,6,7,8-HpCDD	$C_{12}HCl_7O_2$
Triazinone	**41394-05-2**	Metamitron	$C_{10}H_{10}N_4O$	PCDD	**39227-28-6**	1,2,3,4,7,8-HxCDD	$C_{12}H_2Cl_6O_2$
Triazole	**61-82-5**	Amitrole	$C_2H_4N_4$	PCDD	**40321-76-4**	1,2,3,7,8-PeCDD	$C_{12}H_3Cl_5O_2$
Triazole	**76674-21-0**	Flutriafol	$C_{16}H_{13}F_2N_3O$	PCDD	**57653-85-7**	1,2,3,6,7,8-HxCDD	$C_{12}H_2Cl_6O_2$
Triazole	**94361-06-5**	Cyproconazole	$C_{15}H_{18}ClN_3O$				
Triazole	**119446-68-3**	Difenoconazole	$C_{19}H_{17}Cl_2N_3O_3$	Medicine	**298-46-4**	Carbamazepine	$C_{15}H_{12}N_2O$
Triazole	**131983-72-7**	Triticonazole	$C_{17}H_{20}ClN_3O$	Medicine	**14168-01-5**	Dilor	$C_{10}H_7Cl_7$
Triazole	**133855-98-8**	Epoxiconazole	$C_{17}H_{13}ClFN_3O$				
Thiocarbamate	**137-26-8**	Thiram	$C_6H_{12}N_2S_4$	Hormone	**50-28-2**	Estradiol	$C_{18}H_{24}O_2$
Thiocarbamate	**759-94-4**	EPTC	$C_9H_{19}NOS$				
Thiocarbamate	**1929-77-7**	Vernolate	$C_{10}H_{21}NOS$	Auxin	**87-51-4**	Indolylacetic acid	$C_{10}H_9NO_2$
Thiocarbamate	**2303-16-4**	Di-allate	$C_{10}H_{17}Cl_2NOS$				
Thiocarbamate	**2303-17-5**	Tri-allate	$C_{10}H_{16}Cl_3NOS$	Other	**608-73-1**	Hexachlorocyclohexane	$C_6H_6Cl_6$
Chloroacetamide	**1918-16-7**	Propachlor	$C_{11}H_{14}ClNO$	Other	**2550-75-6**	Chlorbicyclene	$C_9H_6Cl_8$
Chloroacetamide	**15972-60-8**	Alachlor	$C_{14}H_{20}ClNO_2$				
Chloroacetamide	**34256-82-1**	Acetochlor	$C_{14}H_{20}ClNO_2$				
Chloroacetamide	**51218-45-2**	Metolachlor	$C_{15}H_{22}ClNO_2$				
Chloroacetamide	**67129-08-2**	Metazachlor	$C_{14}H_{16}ClN_3O$				
Dinitroaniline	**1582-09-8**	Trifluralin	$C_{13}H_{16}F_3N_3O_4$				
Dinitroaniline	**19044-88-3**	Oryzalin	$C_{12}H_{18}N_4O_6S$				
Dinitroaniline	**33629-47-9**	Butralin	$C_{14}H_{21}N_3O_4$				
Dinitroaniline	**40487-42-1**	Pendimethalin	$C_{13}H_{19}N_3O_4$				
Pyrethroid	**52315-07-8**	Cypermethrin	$C_{22}H_{19}Cl_2NO_3$				
Pyrethroid	**52645-53-1**	Permethrin	$C_{21}H_{20}Cl_2NO_3$				
Pyrethroid	**68359-37-5**	Cyfluthrin	$C_{22}H_{18}Cl_2FNO_3$				
Triketone	**99105-77-8**	Sulcotrione	$C_{14}H_{13}ClO_5S$				
Triketone	**335104-84-2**	Tembotrione	$C_{17}H_{16}ClF_3O_6S$				
Phthalimide	**133-06-2**	Captan	$C_9H_8Cl_3NO_2S$				
Phthalimide	**133-07-3**	Folpet	$C_9H_4Cl_3NO_2S$				
Cyclodiene	**309-00-2**	Aldrine	$C_{12}H_8Cl_6$				
Cyclodiene	**465-73-6**	Isodrine	$C_{12}H_8Cl_6$				
Aryloxyalkanoic acid	**94-74-6**	MCPA	$C_9H_9ClO_3$				
Aryloxyalkanoic acid	**7085-19-0**	Mecoprop	$C_{10}H_{11}ClO_3$				
Alkylchlorophenoxy	**94-75-7**	2,4-D	$C_8H_6Cl_2O_3$				
Phosphonoglycine	**1071-83-6**	Glyphosate	$C_3H_7NO_5P$				
Chloronitrile	**1897-45-6**	Chlorothalonil	$C_8Cl_4N_2$				
Benzoic acid	**1918-00-9**	Dicamba	$C_8H_6Cl_2O_3$				
Pyridine	**1918-02-1**	Picloram	$C_6H_3Cl_3N_2O_2$				
Sulfite ester	**2312-35-8**	Propargite	$C_{19}H_{26}O_4S$				
Ethylene generator	**16672-87-0**	Ethephon	$C_2H_6ClO_3P$				
Dicarboximide	**36734-19-7**	Iprodione	$C_{13}H_{13}Cl_2N_3O_3$				
Aryloxyphenoxypropionate	**51338-27-3**	Diclofop-methyl	$C_{16}H_{14}Cl_2O_4$				
Diphenyl ether	**74070-46-5**	Aclonifen	$C_{12}H_9ClN_2O_3$				
Anilinopyrimidine	**121552-61-2**	Cyprodinil	$C_{14}H_{15}N_3$				

Hydroxyanilide	**126833-17-8**	Fenhexamid	$C_{14}H_{17}Cl_2NO_2$
Neonicotinoid	**135410-20-7**	Acetamiprid	$C_{10}H_{11}ClN_4$
Diphenyl oxazoline	**153233-91-1**	Etoxazole	$C_{21}H_{23}F_2NO_2$

Table A2

List of degradation products inputted in TyPol. When there was no referenced CAS, a number was created according to the following format: CAS number of the parent-INRA-i (i represents the number of the degradation product among all degradation products of the parent pompound)

CAS number	Chemical formula	Parent CAS number	Parent name
64-19-7	CH_3COOH	34256-82-1	Acetochlor
72-54-8	$C_{14}H_{10}Cl_4$	50-29-3	p,p'-DDT
72-55-9	$C_{14}H_8Cl_4$	50-29-3	p,p'-DDT
85-41-6	$C_8H_5NO_2$	133-07-3	Folpet
88-97-1	$C_8H_7NO_3$	133-07-3; 131983-72-7	Folpet; Triticonazole
88-99-3	$C_6H_4(COOH)_2$	133-07-3	Folpet
95-76-1	$C_6H_5Cl_2N$	330-54-1	Diuron
1024-57-3	$C_{10}H_5Cl_7O$	57-74-9; 76-44-8	Chlordane; Heptachlore
1031-07-8	$C_9H_6Cl_6O_4S$	115-29-7	Endosulfan
1570-64-5	C_7H_7ClO	94-74-6	MCPA
1897-45-6-INRA-1	$C_8HCl_3N_2O$	1897-45-6	Chlorothalonil
1897-45-6-INRA-2	$C_8H_3Cl_3N_2O_4S$	1897-45-6	Chlorothalonil
2327-02-8	$C_7H_6Cl_2N_2O$	330-54-1	Diuron
3567-62-2	$C_8H_8Cl_2N_2O$	330-54-1	Diuron
3739-38-6	$C_{13}H_{10}O_3$	52315-07-8	Cypermethrin
27304-13-8	$C_{10}H_4Cl_8O$	57-74-9	Chlordane
34256-82-1-INRA-1	$C_{14}H_{21}NO_5S$	34256-82-1	Acetochlor
34256-82-1-INRA-2	$C_{14}H_{19}NO_4$	34256-82-1	Acetochlor
34256-82-1-INRA-3	$C_{16}H_{23}NO_5S$	34256-82-1	Acetochlor
63637-89-8	$C_{17}H_{20}ClN_3O_2$	36734-19-7	Iprodione
68359-37-5-INRA-1	$C_8H_{10}Cl_2O_2$	68359-37-5	Cyfluthrin
77279-89-1	$C_{13}H_9FO_3$	68359-37-5	Cyfluthrin
131983-72-7-INRA-1	$C_{17}H_{20}ClN_3O_2$	131983-72-7	Triticonazole
131983-72-7-INRA-2	$C_{17}H_{20}ClN_3O_2$	131983-72-7	Triticonazole

Table A3

Example of TyPol results: clustering of the 215 organic compounds (parent substances and degradation products) inputted in the database (chemical families, CAS numbers, names) (PCB: polychlorinated biphenyls, PAH: polycyclic aromatic hydrocarbons, PCDF: polychlorinated dibenzofurans, PCDD: polychlorinated dibenzodioxins)

Chemical family	CAS number	Name
Cluster 1		
Carbamate	63-25-2	Carbaryl
Carbamate	101-21-3	Chlorpropham
Carbamate	116-06-3	Aldicarb
Carbamate	3337-71-1	Asulam
Carbamate	23135-22-0	Oxamyl
Carbamate	24579-73-5	Propamocarb
Thiocarbamate	137-26-8	Thiram
Thiocarbamate	759-94-4	EPTC
Thiocarbamate	1929-77-7	Vernolate
Thiocarbamate	2303-16-4	Di-allate
Thiocarbamate	2303-17-5	Tri-allate
Urea	330-54-1	Diuron
Urea	330-55-2	Linuron
Urea	1746-81-2	Monolinuron
Urea	3060-89-7	Metobromuron
Urea	15545-48-9	Chlortoluron
Organophosphorous	60-51-5	Dimethoate
Organophosphorous	5598-13-0	Chlorpyrifos-methyl
Organophosphorous	13194-48-4	Ethoprophos
Organophosphorous	15845-66-6	Fosetyl
Triazine	122-34-9	Simazine
Triazine	1912-24-9	Atrazine
Triazine	66215-27-8	Cyromazine
Alkylchlorophenoxy	94-75-7	2,4-D
Amide	709-98-8	Propanil
Aryloxyalkanoic acid	7085-19-0	Mecoprop
Aryloxyalkanoic acid	94-74-6	MCPA
Auxin	87-51-4	Indolylacetic acid
Benzoic acid	1918-00-9	Dicamba
Chloroacetamide	1918-16-7	Propachlor
Chloronitrile	1897-45-6	Chlorothalonil
Ethylene generator	16672-87-0	Ethephon
Neonicotinoid	135410-20-7	Acetamiprid
Phosphonoglycine	1071-83-6	Glyphosate
Phthalates	131-11-3	Dimethyl
Pyridine	1918-02-1	Picloram
Triazinone	21087-64-9	Metribuzin
Triazole	61-82-5	Amitrole
Metabolite	64-19-7	Metabolite of acetochlor
Metabolite	85-41-6	Metabolite of folpet
Metabolite	88-97-1	Metabolite of folpet and triticonazole
Metabolite	88-99-3	Metabolite of folpet
Metabolite	95-76-1	Metabolite of diuron

Metabolite	1570-64-5	Metabolite of MCPA
Metabolite	1897-45-6-INRA-1	Metabolite of chlorothalonil
Metabolite	2327-02-8	Metabolite of diuron
Metabolite	3567-62-2	Metabolite of diuron
Metabolite	68359-37-5-INRA-1	Metabolite of cyfluthrin
Cluster 2		
PCB	92-52-4	Biphenyl
PCB	2050-67-1	3,3'-dichlorobiphenyl
PCB	2050-68-2	4,4'-dichlorobiphenyl
PCB	2051-60-7	2-chlorobiphenyl
PCB	2051-61-8	3-chlorobiphenyl
PCB	2974-90-5	3,4'-dichlorobiphenyl
PCB	2974-92-7	3,4-dichlorobiphenyl
PCB	13029-08-8	2,2'-dichlorobiphenyl
PCB	16605-91-7	2,3-dichlorobiphenyl
PCB	25569-80-6	2,3'-dichlorobiphenyl
PCB	33146-45-1	2,6-dichlorobiphenyl
PCB	33284-50-3	2,4-dichlorobiphenyl
PCB	34883-39-1	2,5-dichlorobiphenyl
PCB	34883-41-5	3,5-dichlorobiphenyl
PCB	34883-43-7	2,4'-dichlorobiphenyl
PCB	37680-66-3	2,2',4-trichlorobiphenyl
PCB	38444-78-9	2,2',3-trichlorobiphenyl
PCB	38444-93-8	2,2',3,3'-tetrachlorobiphenyl
PCB	52663-59-9	2,2',3,4-tetrachlorobiphenyl
PCB	52663-62-4	2,2',3,3',4-pentachlorobiphenyl
PCB	60145-20-2	2,2',3,3',5-pentachlorobiphenyl
PAH	85-01-8	Phenanthrene
PAH	86-73-7	Fluorene
PAH	120-12-7	Anthracene
Organochlorine	58-89-9	Lindane
Organochlorine	118-74-1	Hexachlorobenzene
Urea	101-42-8	Fenuron
Other	608-73-1	Hexachlorocyclohexane
Metabolite	72-54-8	Metabolite of p-p'-DDT
Metabolite	72-55-9	Metabolite of p-p'-DDT
Cluster 3		
PAH	50-32-8	Benzo(a)pyrene
PAH	53-70-3	Dibenzo(a,h)anthracene
PAH	56-55-3	Benzo(a)anthracene
PAH	129-00-0	Pyrene
PAH	191-24-2	Benzo(g,h,i)perylene
PAH	193-39-5	Indeno[1,2,3-cd]pyrene
PAH	205-99-2	Benzo(b)fluoranthene
PAH	206-44-0	Fluoranthene
PAH	207-08-9	Benzo(k)fluoranthene
PAH	218-01-9	Chrysene
PCDF	39001-02-0	OCDF
PCDF	51207-31-9	2,3,7,8-TCDF
PCDF	55673-89-7	1,2,3,4,7,8,9-HpCDF
PCDF	57117-31-4	2,3,4,7,8-PeCDF
PCDF	57117-41-6	1,2,3,7,8-PeCDF
PCDF	57117-44-9	1,2,3,6,7,8-HxCDF
PCDF	60851-34-5	2,3,4,6,7,8-HxCDF
PCDF	67562-39-4	1,2,3,4,6,7,8-HpCDF
PCDF	70648-26-9	1,2,3,4,7,8-HxCDF

PCDF	72918-21-9	1,2,3,7,8,9-HxCDF
Organochlorine	50-29-3	p,p'-DDT
Organochlorine	57-74-9	Chlordane
Organochlorine	60-57-1	Dieldrine
Organochlorine	72-20-8	Endrine
Organochlorine	76-44-8	Heptachlore
Organochlorine	115-29-7	Endosulfan
Organochlorine	297-78-9	Isobenzane
Organochlorine	1715-40-8	Bromocyclene
Organochlorine	8001-35-2	Toxaphene
PCB	2051-24-3	Decachlorobiphenyl
PCB	35065-30-6	2,2',3,3',4,4',5-heptachlorobiphenyl
PCB	35694-08-7	2,2',3,3',4,4',5,5'-octachlorobiphenyl
PCB	38380-07-3	2,2',3,3',4,4'-hexachlorobiphenyl
PCB	40186-72-9	2,2',3,3',4,4',5,5',6-nonachlorobiphenyl
PCB	52663-71-5	2,2',3,3',4,4',6-heptachlorobiphenyl
PCB	52663-78-2	2,2',3,3',4,4',5,6-octachlorobiphenyl
PCB	52663-79-3	2,2',3,3',4,4',5,6,6'-nonachlorobiphenyl
PCB	55215-18-4	2,2',3,3',4,5-hexachlorobiphenyl
PCDD	1746-01-6	2,3,7,8-tetrachloro-dibenzo-p-dioxine
PCDD	3268-87-9	OCDD
PCDD	19408-74-3	1,2,3,7,8,9-HxCDD
PCDD	35822-46-9	1,2,3,4,6,7,8-HpCDD
PCDD	39227-28-6	1,2,3,4,7,8-HxCDD
PCDD	40321-76-4	1,2,3,7,8-PeCDD
PCDD	57653-85-7	1,2,3,6,7,8-HxCDD
Cyclodiene	309-00-2	Aldrine
Cyclodiene	465-73-6	Isodrine
Medicine	14168-01-5	Dilor
Other	2550-75-6	Chlorbicyclene
Metabolite	1024-57-3	Metabolite of chlordane and heptachlore
Metabolite	1031-07-8	Metabolite of endosulfan
Metabolite	27304-13-8	Metabolite of chlordane
Cluster 4		
Strobilurin	117428-22-5	Picoxystrobin
Strobilurin	131860-33-8	Azoxystrobin
Strobilurin	141517-21-7	Trifloxystrobin
Strobilurin	143390-89-0	Kresoxim-methyl
Strobilurin	149961-52-4	Dimoxystrobin
Strobilurin	175013-18-0	Pyraclostrobin
Strobilurin	361377-29-9	Fluoxastrobin
Phthalates	84-61-7	Di-cyclohexyl
Phthalates	85-68-7	Benzylbutyl
Phthalates	117-81-7	Di-2-ethylhexyl
Phthalates	117-84-0	Di-n-octyl
Phthalates	26761-40-0	Di-isodecyl
Phthalates	28553-12-0	Di-isononyl
Triazole	76674-21-0	Flutriafol
Triazole	94361-06-5	Cyproconazole
Triazole	119446-68-3	Difenoconazole
Triazole	131983-72-7	Triticonazole
Triazole	133855-98-8	Epoxiconazole
Amide	15299-99-7	Napropamide
Amide	55814-41-0	Mepronil
Amide	180409-60-3	Cyflufenamid
Pyrethroid	52315-07-8	Cypermethrin
Pyrethroid	52645-53-1	Permethrin
Pyrethroid	68359-37-5	Cyfluthrin

Carbamate	13684-56-5	Desmedipham
Carbamate	13684-63-4	Phenmedipham
Carbamate	79127-80-3	Fenoxycarb
Aryloxyphenoxypropionate	51338-27-3	Diclofop-methyl
Dicarboximide	36734-19-7	Iprodione
Diphenyl oxazoline	153233-91-1	Etoxazole
Hormone	50-28-2	Estradiol
Hydroxyanilide	126833-17-8	Fenhexamid
Sulfite ester	2312-35-8	Propargite
Triketone	335104-84-2	Tembotrione
Metabolite	131983-72-7-INRA-1	Metabolite of triticonazole
Metabolite	131983-72-7-INRA-2	Metabolite of triticonazole
Metabolite	63637-89-8	Metabolite of iprodione
Cluster 5		
Organophosphorous	56-38-2	Parathion-ethyl
Organophosphorous	86-50-0	Azinphos-methyl
Organophosphorous	121-75-5	Malathion
Organophosphorous	122-14-5	Fenitrothion
Organophosphorous	333-41-5	Diazinon
Amide	23950-58-5	Propyzamide
Amide	35256-85-0	Tebutam
Amide	57837-19-1	Metalaxyl
Amide	77732-09-3	Oxadixyl
Chloroacetamide	15972-60-8	Alachlor
Chloroacetamide	34256-82-1	Acetochlor
Chloroacetamide	51218-45-2	Metolachlor
Chloroacetamide	67129-08-2	Metazachlor
Dinitroaniline	1582-09-8	Trifluralin
Dinitroaniline	19044-88-3	Oryzalin
Dinitroaniline	33629-47-9	Butralin
Dinitroaniline	40487-42-1	Pendimethalin
Triazine	834-12-8	Ametryn
Triazine	886-50-0	Terbutryn
Triazine	5915-41-3	Terbuthylazine
Triazine	21725-46-2	Cyanazine
Urea	555-37-3	Neburon
Urea	34123-59-6	Isoproturon
Urea	64902-72-3	Chlorsulfuron
Urea	79510-48-8	Metsulfuron
Carbamate	1563-66-2	Carbofuran
Carbamate	16118-49-3	Carbetamide
Carbamate	23103-98-2	Pirimicarb
Diazine	1698-60-8	Chloridazon
Diazine	25057-89-0	Bentazone
Phthalate	84-66-2	Diethyl
Phthalate	84-74-2	Dibutyl
Phthalimide	133-06-2	Captan
Phthalimide	133-07-3	Folpet
Diphenyl ether	74070-46-5	Aclonifen
Medicine	298-46-4	Carbamazepine
Anilinopyrimidine	121552-61-2	Cyprodinil
Triazinone	41394-05-2	Metamitron
Triketone	99105-77-8	Sulcotrione
Metabolite	1897-45-6-INRA-2	Metabolite of chlorothalonil
Metabolite	3739-38-6	Metabolite of cypermethrin
Metabolite	34256-82-1-INRA-1	Metabolite of acetochlor
Metabolite	34256-82-1-INRA-2	Metabolite of acetochlor
Metabolite	34256-82-1-INRA-3	Metabolite of acetochlor

Metabolite	77279-89-1	Metabolite of cyfluthrin
Cluster 6		
Organochlorine	143-50-0	Chlordecone
Organochlorine	2385-85-5	Mirex
Organochlorine	4234-79-1	Kelevan