



HAL
open science

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

► To cite this version:

Rémi Servien, Laure Mamy, Ziang Li, Virginie Rossard, Eric Latrille, et al.. TyPol - a new methodology for organic compounds clustering based on their molecular characteristics and environmental behavior. *Chemosphere*, 2014, 111, pp.613-622. 10.1016/j.chemosphere.2014.05.020 . hal-00924015v2

HAL Id: hal-00924015

<https://hal.science/hal-00924015v2>

Submitted on 12 May 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3
4 Rémi Servien^{a,b,*}, Laure Mamy^c, Ziang Li^d, Virginie Rossard^b, Eric Latrille^b, Fabienne
5 Bessac^{e,f,g}, Dominique Patureau^b, Pierre Benoit^d

6
7 ^a INRA, Université de Toulouse, UMR 1331 Toxalim, Research Centre in Food Toxicology,
8 F-31027 Toulouse, France (present address)

9 ^b INRA, UR 050, Laboratoire de Biotechnologie de l'Environnement, Avenue des Etangs, F-
10 11100 Narbonne, France

11 ^c INRA, UR 251 PESSAC, Route de St Cyr, F-78026 Versailles, France

12 ^d UMR 1091 INRA-AgroParisTech, Environnement et Grandes Cultures, F-78850 Thiverval-
13 Grignon, France

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

16 ^f Université de Toulouse, UPS, IRSAMC, Laboratoire de Chimie et Physique Quantiques, 118
17 route de Narbonne, F-31062 Toulouse, France

18 ^g CNRS (UMR 5626), F-31062 Toulouse, France

19

20 * Corresponding author. Address: INRA, Université de Toulouse, UMR 1331 Toxalim, Research
21 Centre in Food Toxicology, F-31027 Toulouse, France. Tel.: +33 (0)5 61 19 32 82; fax: +33 (0)5
22 61 19 39 17.

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

1 HIGHLIGHTS

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

8 9 ABSTRACT

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

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

6

7 *Keywords:*

8 Pesticides

9 Degradation products

10 Clustering

11 Molecular modeling

12 Environmental fate

13 Partial least squares

14

15 **1. Introduction**

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

1 environment and that originate from various sources. According to authors, from 30 000 to 100
2 000 chemical substances may be concerned by environmental risks assessment (Muir and
3 Howard, 2006). However, their behavior in the environment and their transfer to environmental
4 components such as water or atmosphere are studied for only a very small proportion of the
5 chemical in laboratory tests or monitoring studies because it is time-consuming and/or cost
6 prohibitive. Consequently, a high number of different in silico approaches have been developed
7 to estimate the behavior of organic compounds in the environment. The most used in silico
8 approaches, that are also the most simple, are based on QSAR (Quantitative Structure Activity
9 Relationship). QSARs allow the estimation of one or several compound properties (such as
10 sorption by soils and sediments, biodegradation, bioconcentration factor or biological activities)
11 from some other properties such as structural molecular properties (number of atoms in the
12 molecule, molecular surface, dipole moment, energy of orbitals...), water solubility or octanol-
13 water partition coefficient (e.g. OECD, 1993a; Raymond et al., 2001; Worrall, 2001; Eriksson et
14 al., 2002; Pavan et al., 2008). Other approaches aim, for example, at ranking organic compounds
15 knowing the values of several of their properties such as partitioning, persistence or
16 bioaccumulation. Compounds that have undesirable properties can be considered for
17 management, regulation, or even global bans on production and use (Mackay et al., 2001;
18 Walker and Carlsen, 2002). Numerical models represent the most complex approaches as they
19 allow overall assessment of the fate of organic compounds in the soil, water and air, and as they
20 take into account the physico-chemical properties of the compounds and the agro-pedo-climatic
21 conditions (e.g. Carsel et al., 1998; Jarvis and Larsbo, 2012). However, they require a lot of input
22 data. Therefore, approaches able to classify compounds according to their environmental
23 behavior or eco/toxicological effects will help both regulators and scientists facing the problem

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

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

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

15 The choice of the statistical method involved in TyPol is crucial for the reliability of the
16 clustering. Principal components analysis (PCA) is often used in multivariate chemical
17 characterizations to determine linearly uncorrelated variables that summarize the information
18 contained in variables (Jackson, 1991; Snarey et al., 1997; Harju et al., 2002; Eriksson et al.,
19 2006). These uncorrelated variables can also be used as an excellent basis to select a
20 representative set of chemicals using clustering methods. Various clustering techniques have
21 been employed in chemical mapping such as strategies based on PCA and hierarchical clustering
22 for selecting dissimilar organic substances (Rännar and Anderson, 2010) or ranking non-ionic
23 organic pesticides (Gramatica et al., 2004), bayesian classifiers for chemical toxicity predictions

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

14

15 **2. Materials and methods**

16 *2.1. Organic compounds*

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

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

8

9 *2.2. Environmental processes and parameters*

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

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

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

13 *2.3. Molecular descriptors: selection and calculation*

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

1 In addition, five criteria were defined to choose the descriptors: (i) their relevance to
2 estimate the seven environmental parameters (see the cited references below), (ii) their common
3 use for the estimation of these seven parameters, (iii) the absence of redundancy between
4 descriptors, (iv) the possibility to calculate the descriptors with molecular modeling, and (v) their
5 ranges of variation. Finally, 40 constitutional, geometrical, topological, and quantum-chemical
6 descriptors were retained (see for example OECD, 1993b; Katritzky et al., 2000; Sabljic, 2001;
7 Dearden and Schüürmann, 2003; Doucette, 2003; Yang et al., 2003; Pavan et al., 2008) (Table
8 2).

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

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

3

4 *2.4. Partial least squares (PLS) regression*

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

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

16

17 *2.5. Domain of validity*

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

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

3

4 *2.6. Hierarchical clustering*

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

16

17 *2.7. Robustness of the method*

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

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

10

11 2.8. *Computing tools*

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

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

3

4 **3. Results and discussion**

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

15

16 *3.1. Chemical mapping by PLS*

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

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

1 1989; ATSDR, 1995; Cabidoche et al., 2009; Dolfig et al., 2012) contrary to fosetyl which has
2 a very low *DT50* (PPDB, 2013); and di-isodecyl, di-isononyl, and benzo(g,h,i)perylene have
3 very high *Kow* values (PPDB, 2013). Chlordecone, benzo(g,h,i)perylene, mirex and kelevan also
4 have very high connectivity indexes. Nevertheless, these compounds were taken into
5 consideration for the subsequent analysis because they could be representative of other
6 compounds.

7 The four-component PLS model has good statistical results: $R^2_X=0.77$, $R^2_Y=0.90$ and
8 $Q^2_Y=0.44$. These results shows that the PLS is a good model for the different compounds
9 included in TyPol. The first two components were the most important ones. The closer the
10 compounds are in this score-plot, the more similar they are (Fig. 2). The main characteristic of
11 the first component, which explains 40% of the variance, is the strong positive loadings for all
12 the geometric and topological descriptors, and constitutional descriptors like the number of
13 chlorine or halogen atoms. A contrario, the dipole moment and the total energy have strong
14 negative loadings therefore have an opposite effect. The second axis explains 16% of the
15 variance. On this axis, variables such as the number of chlorine or halogen atoms have a positive
16 loading whereas the number of rotatable, double or simple bonds or the number of hydrogen,
17 oxygen or total atoms have a negative loading (Fig. 3). Figure 3 also shows that many variables
18 seem to be correlated, mainly the different connectivity and valence connectivity indexes. On the
19 third axis, variables such as the number of carbon or hydrogen atoms, and the molecular weight
20 have strong positive loadings, and others such as the number of circuits or the LUMO energy are
21 on the opposite side. On the fourth axis, the HOMO energy and the numbers of rings atoms are
22 on opposite sides of the number of sulfur or chlorine atoms.

23

1 3.2. Clustering

2 Using a hierarchical clustering algorithm, several clusterings, from 1 (all compounds in
3 the same cluster) to 215 (all compounds in a different cluster), were obtained. The selection of
4 the number of clusters is an important and difficult task, which is usually performed by plotting
5 the heights of the dendrogram's node and looking for a break. The results showed that the best
6 choice was to classify the compounds in 6 clusters. The size of the six clusters varied from 3 to
7 52 compounds (Fig. 2, Table A3), each cluster being characterized by specific features.

8 Figure 4 shows the range of variations of the values of the 7 environmental parameters
9 for each of the 6 clusters. The importance of the different parameters can be evaluated in Figure
10 3, but Figure 4 provides a description of the characteristics of each cluster. The trigger value of
11 P_{vap} is indicated to differentiate volatile and non-volatile compounds (there is no trigger value for
12 K_H , FOCUS, 2008), that of K_{oc} to differentiate mobile and non-mobile compounds (McCall et
13 al., 1980), that of $DT50$ to differentiate persistent and non-persistent compounds (1107/2009 EC,
14 2009), and that of BCF to differentiate compounds having or not a potential of bioaccumulation
15 (1107/2009 EC, 2009) (Fig. 4). Depending on the values of P_{vap} , K_{oc} , $DT50$ and/or BCF , the six
16 clusters aggregate compounds having (or not) risks of air, water and/or soil contamination and/or
17 high ecotoxicity.

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

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

6 Twenty-one of the 30 compounds of cluster 2 are PCB (over 31 inputted in TyPol). There
7 are also 4 organochlorines (2 parent substances and 2 degradation products) and 3 PAH.
8 Compounds of cluster 2 have low dipole moment and high total energy. They also have the
9 lowest $DT50$ of the 6 clusters (rapid dissipation), low K_{oc} (high mobility), but contrary to cluster
10 1, low S_w (and high K_{ow}), and high BCF so high ecotoxicity (Fig. 4).

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

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

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

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

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

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

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

12

13 *3.3. Parents-degradation products relationships*

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

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

8

9 **4. Conclusion**

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

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

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

5
6 **Acknowledgements**

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

13
14 **Appendix A. Supplementary material**

15 Supplementary Tables A1, A2, A3

16
17 **References**

18 Ahlers, J., Stock, F., Werschkun, B., 2008. Integrated testing and intelligent assessment-new
19 challenges under REACH. *Environ. Sci. Pollut. Res.* 15, 565-572.
20 ATSDR (Agency for Toxic Substances and Disease Registry), 1995. Toxicological profile for
21 mirex and chlordecone. US Department of Health and Human Services, Public Health
22 Services, <http://www.atsdr.cdc.gov/toxprofiles/tp66.pdf>

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

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

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

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

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

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

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

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

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

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

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

8 Gramatica, P., Papa, E., Battaini, F., 2004. Ranking and classification of non-ionic organic
9 pesticides for environmental distribution: a QSAR approach. *Int. J. Env. Anal. Chem.* 84,
10 65-74.

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

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

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

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

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

21 Katritzky, A.R., Maran, U., Lobanov, V.S., Karelson, M., 2000. Structurally diverse quantitative
22 structure-property relationship correlations of technologically relevant physical properties.
23 *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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

8 Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*
9 58, 236-244.

10 Wold, H., 1996. Estimation of principal component and related models by iterative least squares,
11 in: Krishnaiah, P.R. (Ed.), *Multivariate Analysis*, Academic Press, New York, pp. 391-
12 420.

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

15 Yang, P., Chen, J., Chen, S., Yuan, X., Schramm, K.-W., Kettrup, A., 2003. QSPR models for
16 physicochemical properties of polychlorinated diphenyl ethers. *Sci. Tot. Environ.* 305, 65-
17 76.

18
19
20
21
22
23
24

1 **Figure captions**

2

3 **Fig. 1.** Construction and structure of TyPol. A “new” compound is either a parent compound or a
4 degradation product, for which the environmental parameters (for example water solubility S_w ,
5 adsorption coefficient K_{oc} or bioconcentration factor BCF) are not known.

6

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

9

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

14

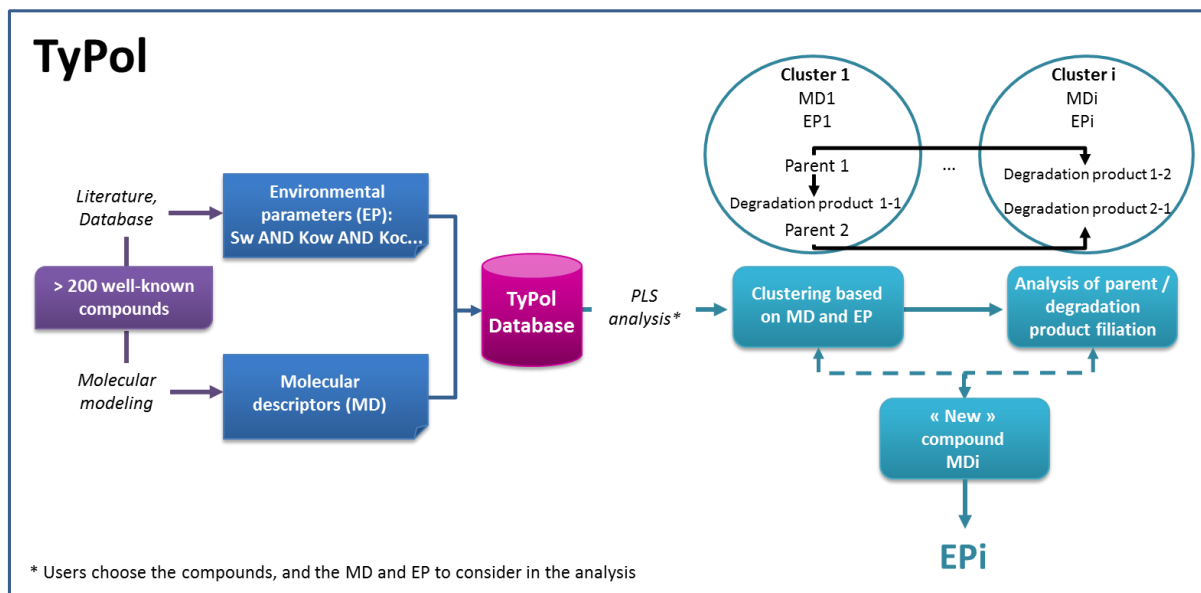
15 **Fig. 4.** Range of variation (box-and-whisker plots) of the values of the seven environmental
16 parameters considered into TyPol (water solubility S_w , octanol-water partition coefficient K_{ow} ,
17 vapor pressure P_{vap} , Henry’s law constant K_H , adsorption coefficient normalized to soil carbon
18 organic content K_{oc} , half-life $DT50$, and bioconcentration factor BCF) for each cluster after
19 analysis of the 215 organic compounds. Dotted lines represent the limits between: volatile (log
20 $P_{vap} > -1$) and non-volatile compounds (log $P_{vap} < -1$) (FOCUS, 2008); mobile (log $K_{oc} < 2.7$)
21 and non-mobile compounds (log $K_{oc} > 2.7$) (McCall et al., 1980); persistent (log $DT50 > 2.25$)
22 and non-persistent compounds (log $DT50 < 2.25$) (Regulation EC 1107/2009, 2009), and
23 compounds having (log $BCF > 2$) or not (log $BCF < 2$) a potential of bioaccumulation
24 (Regulation EC 1107/2009, 2009).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

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.

1 Fig. 1.

2



3

4

5

6

7

8

9

10

11

12

13

14

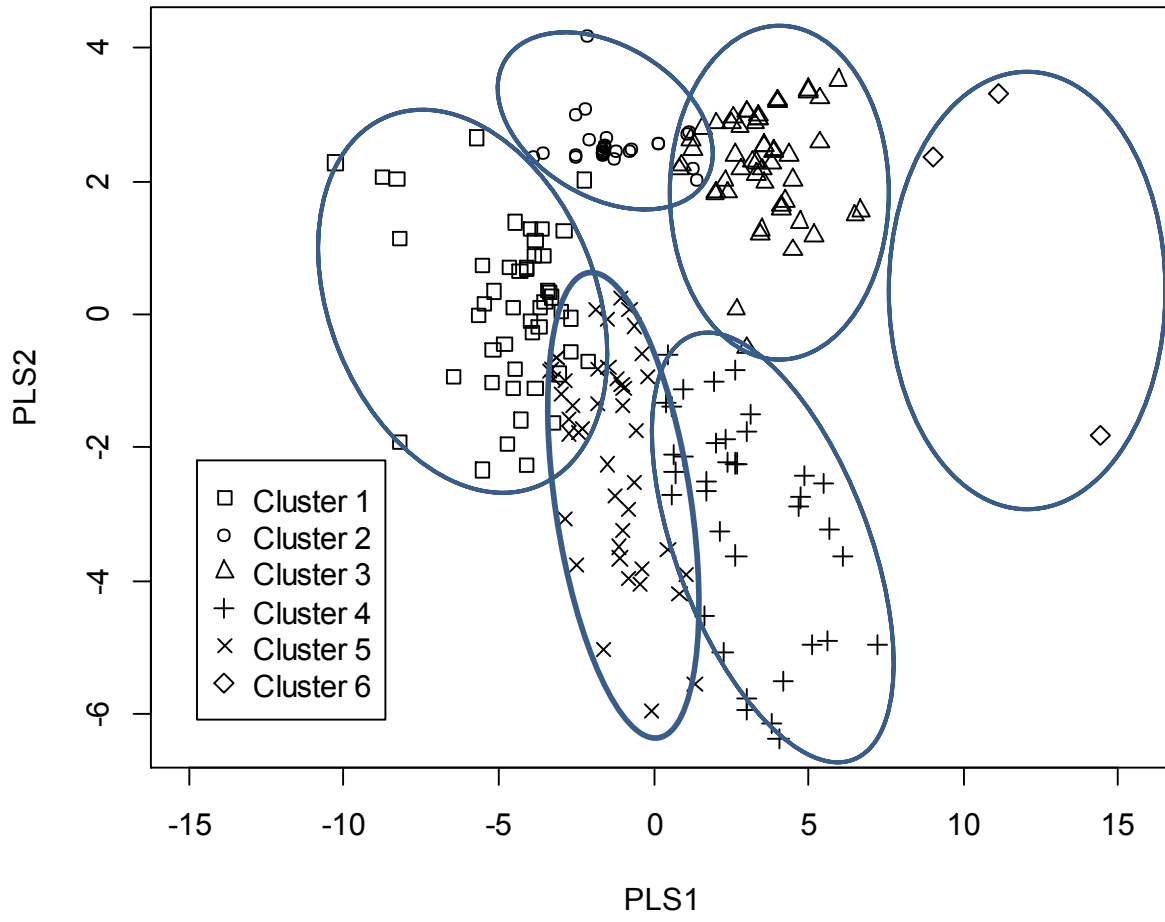
15

16

17

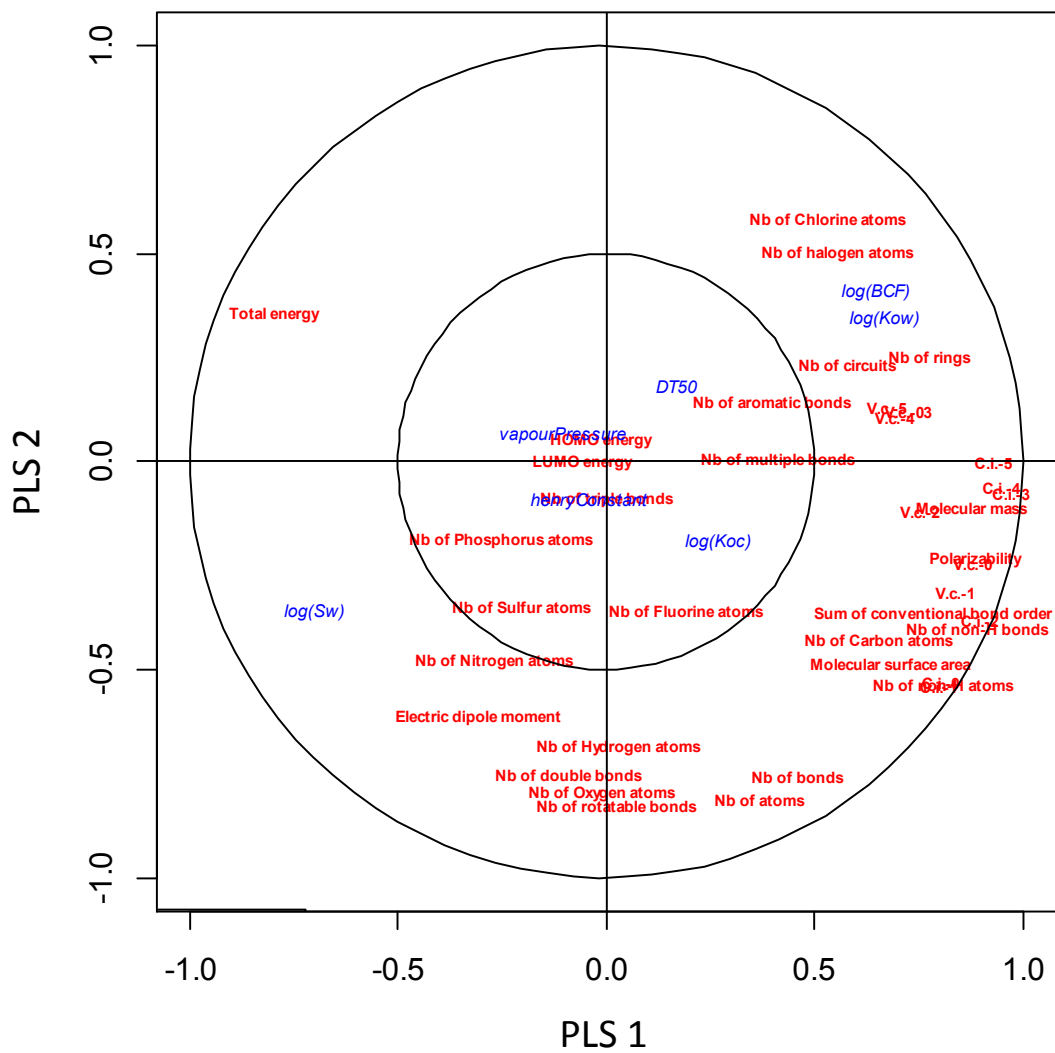
18

1 Fig. 2.



2
3
4
5
6
7
8
9

1 Fig. 3.



2

3

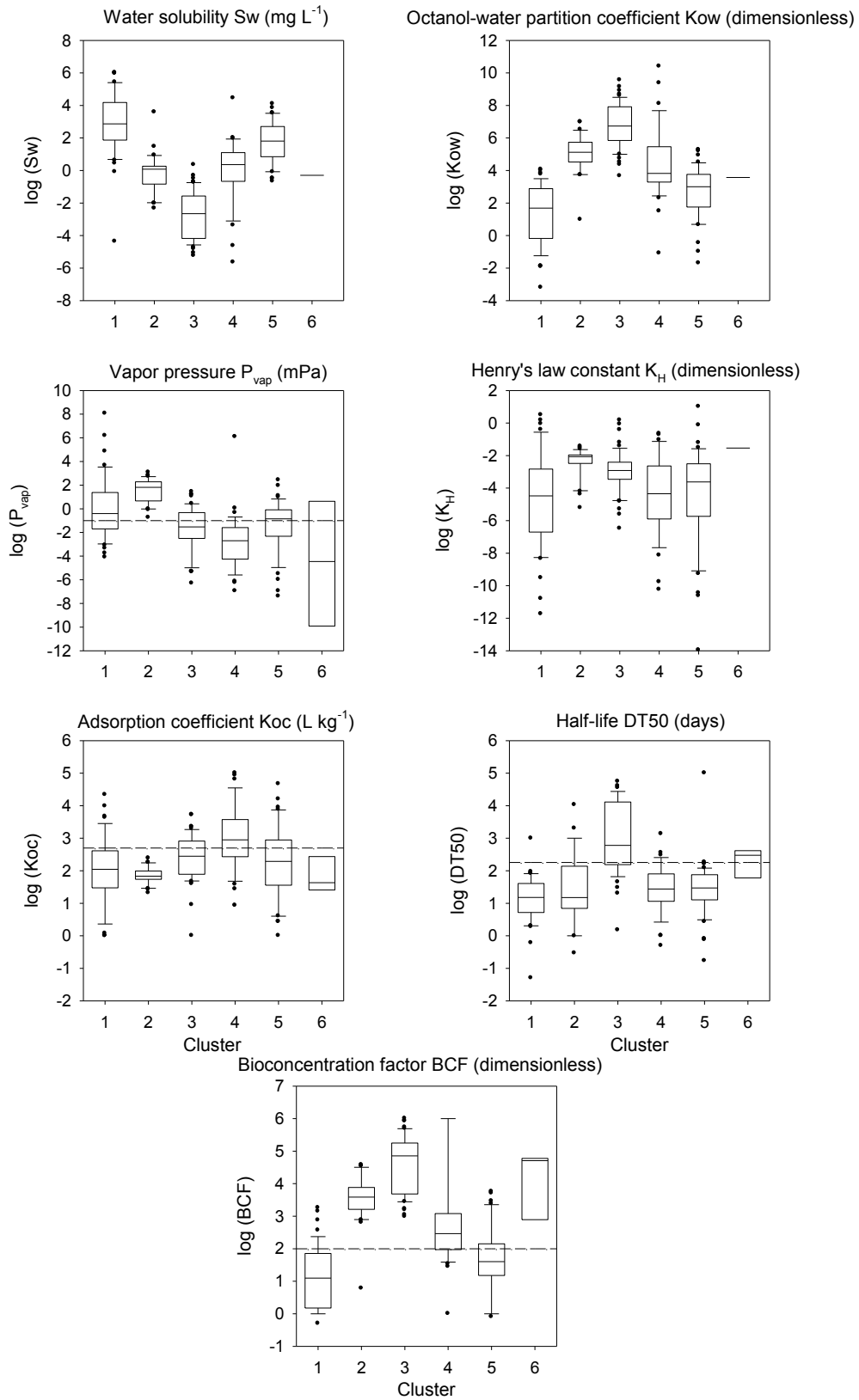
4

5

6

7

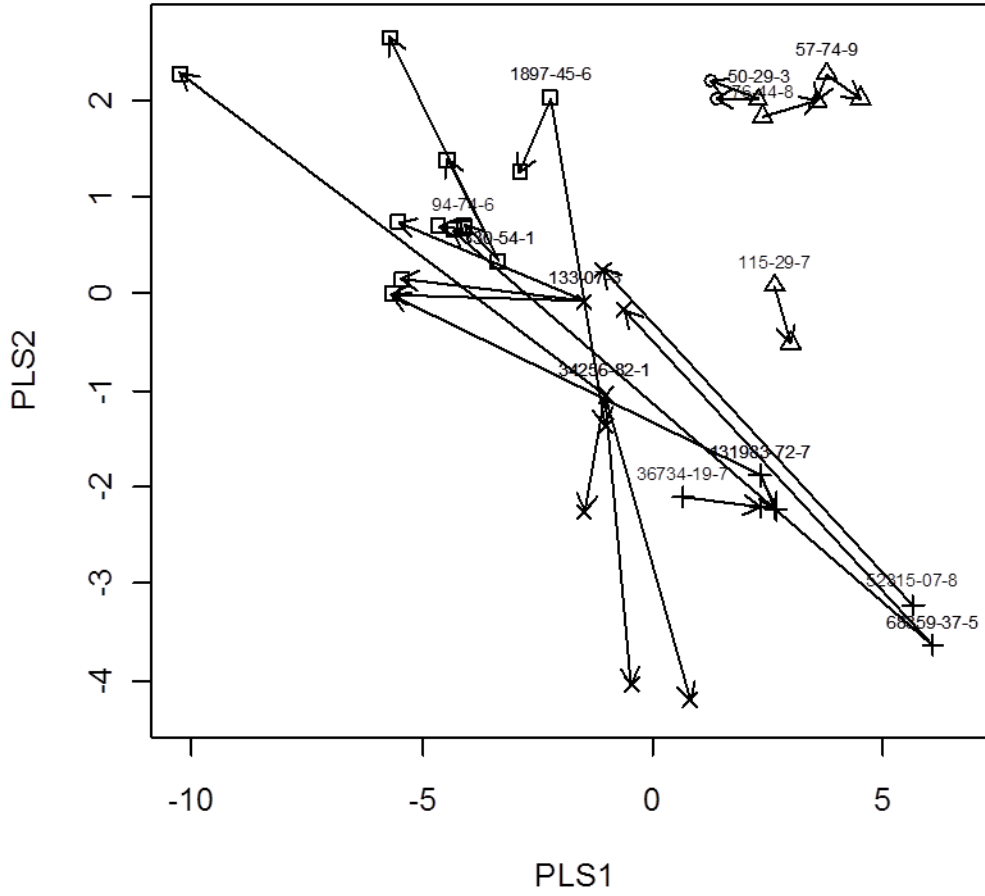
1 **Fig. 4.**



2

1 **Fig. 5.**

2



3

4

5

6

7

8

9

10

11

1 **Tables**

2 **Table 1**

3 Environmental parameters and ranges of variation of their values for the 215 organic compounds
 4 (191 parent compounds and 24 degradation products) inputted in TyPol (*S_w*: water solubility,
 5 *K_{ow}*: octanol-water partition coefficient, *P_{vap}*: vapor pressure, *K_H*: Henry's law constant, *K_{oc}*:
 6 adsorption coefficient, *DT50*: half-life, *BCF*: bioconcentration factor)

7

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

8

9

10

11

12

13

1 **Table 2**

2 List of the 40 molecular descriptors inputted in TyPol and ranges of variation of their values for
 3 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] |

4

1 **Supplementary material**

2 **Table A1**

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

5

| 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 ₉ Cl ₅ | PCB | 92-52-4 | Biphenyl | C ₁₂ H ₁₀ |
| Organochlorine | 57-74-9 | Chlordane | C ₁₀ H ₆ Cl ₈ | PCB | 2050-67-1 | 3,3'-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 58-89-9 | Lindane | C ₆ H ₆ Cl ₆ | PCB | 2050-68-2 | 4,4'-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 60-57-1 | Dieldrine | C ₁₂ H ₈ Cl ₆ O | PCB | 2051-24-3 | Decachlorobiphenyl | C ₁₂ Cl ₁₀ |
| Organochlorine | 72-20-8 | Endrine | C ₁₂ H ₈ Cl ₆ O | PCB | 2051-60-7 | 2-chlorobiphenyl | C ₁₂ H ₉ Cl |
| Organochlorine | 76-44-8 | Heptachlore | C ₁₀ H ₅ Cl ₇ | PCB | 2051-61-8 | 3-chlorobiphenyl | C ₁₂ H ₉ Cl |
| Organochlorine | 115-29-7 | Endosulfan | C ₉ H ₆ Cl ₆ O ₃ S | PCB | 2974-90-5 | 3,4'-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 118-74-1 | Hexachlorobenzene | C ₆ Cl ₆ | PCB | 2974-92-7 | 3,4-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 143-50-0 | Chlordecone | C ₁₀ Cl ₁₀ O | PCB | 13029-08-8 | 2,2'-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 297-78-9 | Isobenzane | C ₉ H ₄ Cl ₈ O | PCB | 16605-91-7 | 2,3-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 1715-40-8 | Bromocyclene | C ₈ H ₅ BrCl ₆ | PCB | 25569-80-6 | 2,3'-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 2385-85-5 | Mirex | C ₁₀ Cl ₁₂ | PCB | 33146-45-1 | 2,6-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 4234-79-1 | Kelevan | C ₁₇ H ₁₂ Cl ₁₀ O ₄ | PCB | 33284-50-3 | 2,4-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Organochlorine | 8001-35-2 | Toxaphene | C ₁₀ H ₈ Cl ₈ | PCB | 34883-39-1 | 2,5-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Carbamate | 63-25-2 | Carbaryl | C ₁₂ H ₁₁ NO ₂ | PCB | 34883-41-5 | 3,5-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Carbamate | 101-21-3 | Chlorpropham | C ₁₀ H ₁₂ ClNO ₂ | PCB | 34883-43-7 | 2,4'-dichlorobiphenyl | C ₁₂ H ₈ Cl ₂ |
| Carbamate | 116-06-3 | Aldicarb | C ₇ H ₁₄ N ₂ O ₂ S | PCB | 35065-30-6 | 2,2',3,3',4,4',5-heptachlorobiphenyl | C ₁₂ H ₃ Cl 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 ₂ Cl ₈ |
| Carbamate | 3337-71-1 | Asulam | C ₈ H ₁₀ N ₂ O ₄ S | PCB | 37680-66-3 | 2,2',4-trichlorobiphenyl | C ₁₂ H ₇ Cl ₃ |
| Carbamate | 13684-56-5 | Desmedipham | C ₁₆ H ₁₆ N ₂ O ₄ | PCB | 38380-07-3 | 2,2',3,3',4,4'-hexachlorobiphenyl | C ₁₂ H ₄ Cl ₆ |
| Carbamate | 13684-63-4 | Phenmedipham | C ₁₆ H ₁₆ N ₂ O ₄ | PCB | 38444-78-9 | 2,2',3-trichlorobiphenyl | C ₁₂ H ₇ Cl ₃ |
| Carbamate | 16118-49-3 | Carbetamide | C ₁₂ H ₁₆ N ₂ O ₃ | PCB | 38444-93-8 | 2,2',3,3'-tetrachlorobiphenyl | C ₁₂ H ₆ Cl ₄ |
| Carbamate | 23103-98-2 | Pirimicarb | C ₁₁ H ₁₈ N ₄ O 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 ₇ H ₁₃ N ₃ O ₃ S | PCB | 52663-59-9 | 2,2',3,4-tetrachlorobiphenyl | C ₁₂ H ₆ Cl ₄ |
| Carbamate | 24579-73-5 | Propamocarb | C ₆ H ₂₀ N ₂ O ₂ | PCB | 52663-62-4 | 2,2',3,3',4-pentachlorobiphenyl | C ₁₂ H ₅ Cl 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 ₃ Cl ₇ |

| | | | | | | | |
|-------------------|--------------------|---------------------|---|-----------|-------------------|--|---|
| Urea | 101-42-8 | Fenuron | C ₉ H ₁₂ N ₂ O | PCB | 52663-78-2 | 2,2',3,3',4,4',5,6-octachlorobiphenyl | C ₁₂ H ₂ Cl ₈ |
| Urea | 330-54-1 | Diuron | C ₉ H ₁₀ Cl ₂ N ₂ O | PCB | 52663-79-3 | 2,2',3,3',4,4',5,6,6'-nonachlorobiphenyl | C ₁₂ HCl ₉ |
| Urea | 330-55-2 | Linuron | C ₉ H ₁₀ Cl ₂ N ₂ O ₂ | PCB | 55215-18-4 | 2,2',3,3',4,5-hexachlorobiphenyl | C ₁₂ H ₄ Cl ₆ |
| Urea | 555-37-3 | Neburon | C ₁₂ H ₁₆ Cl ₂ N ₂ O | PCB | 60145-20-2 | 2,2',3,3',5-pentachlorobiphenyl | C ₁₂ H ₅ Cl ₅ |
| Urea | 1746-81-2 | Monolinuron | C ₉ H ₁₁ ClN ₂ O ₂ | | | | |
| Urea | 3060-89-7 | Metobromuron | C ₉ H ₁₁ BrN ₂ O ₂ | PAH | 50-32-8 | Benzo(a)pyrene | C ₂₀ H ₁₂ |
| Urea | 15545-48-9 | Chlortoluron | C ₁₀ H ₁₃ ClN ₂ O | PAH | 53-70-3 | Dibenzo(a,h)anthracene | C ₂₂ H ₁₄ |
| Urea | 34123-59-6 | Isoproturon | C ₁₂ H ₁₈ N ₂ O | PAH | 56-55-3 | Benzo(a)anthracene | C ₁₈ H ₁₂ |
| Urea | 64902-72-3 | Chlorsulfuron | C ₁₂ H ₁₂ ClN ₅ O ₄ S | PAH | 85-01-8 | Phenanthrene | C ₁₄ H ₁₀ |
| Urea | 79510-48-8 | Metsulfuron | C ₁₃ H ₁₃ N ₅ O ₆ S | PAH | 86-73-7 | Fluorene | C ₁₃ H ₁₀ |
| Organophosphorous | 56-38-2 | Parathion-ethyl | C ₁₀ H ₁₄ NO ₃ PS | PAH | 120-12-7 | Anthracene | C ₁₄ H ₁₀ |
| Organophosphorous | 60-51-5 | Dimethoate | C ₅ H ₁₂ NO ₃ PS ₂ | PAH | 129-00-0 | Pyrene | C ₁₆ H ₁₀ |
| Organophosphorous | 86-50-0 | Azinphos-methyl | C ₁₀ H ₁₂ N ₃ O ₃ PS ₂ | PAH | 191-24-2 | Benzo(g,h,i)perylene | C ₂₂ H ₁₂ |
| Organophosphorous | 121-75-5 | Malathion | C ₁₀ H ₁₉ O ₆ PS ₂ | PAH | 193-39-5 | Indeno[1,2,3-cd]pyrene | C ₂₂ H ₁₂ |
| Organophosphorous | 122-14-5 | Fenitrothion | C ₆ H ₁₂ NO ₅ PS | PAH | 205-99-2 | Benzo(b)fluoranthene | C ₂₀ H ₁₂ |
| Organophosphorous | 333-41-5 | Diazinon | C ₁₂ H ₂₁ N ₂ O ₃ PS | PAH | 206-44-0 | Fluoranthene | C ₁₆ H ₁₀ |
| Organophosphorous | 5598-13-0 | Chlorpyrifos-methyl | C ₇ H ₇ Cl ₃ NO ₃ PS | PAH | 207-08-9 | Benzo(k)fluoranthene | C ₂₀ H ₁₂ |
| Organophosphorous | 13194-48-4 | Ethoprophos | C ₈ H ₁₉ O ₂ PS ₂ | PAH | 218-01-9 | Chrysene | C ₁₈ H ₁₂ |
| Organophosphorous | 15845-66-6 | Fosetyl | C ₂ H ₇ O ₃ P | | | | |
| Amide | 709-98-8 | Propanil | C ₉ H ₉ Cl ₂ NO | PCDF | 39001-02-0 | OCDF | C ₁₂ Cl ₈ O |
| Amide | 15299-99-7 | Napropamide | C ₁₇ H ₂₁ NO ₂ | PCDF | 51207-31-9 | 2,3,7,8-TCDF | C ₁₂ H ₄ Cl ₄ O |
| Amide | 23950-58-5 | Propyzamide | C ₁₂ H ₁₁ Cl ₂ NO | PCDF | 55673-89-7 | 1,2,3,4,7,8,9-HpCDF | C ₁₂ HCl ₇ O |
| Amide | 35256-85-0 | Tebutam | C ₁₅ H ₂₃ NO | PCDF | 57117-31-4 | 2,3,4,7,8-PeCDF | C ₁₂ H ₃ Cl ₅ O |
| Amide | 55814-41-0 | Mepronil | C ₁₇ H ₁₉ NO ₂ | PCDF | 57117-41-6 | 1,2,3,7,8-PeCDF | C ₁₂ H ₃ Cl ₅ O |
| 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 ₂ Cl ₆ O |
| Amide | 77732-09-3 | Oxadixyl | C ₁₄ H ₁₈ N ₂ O ₄ | PCDF | 60851-34-5 | 2,3,4,6,7,8-HxCDF | C ₁₂ H ₂ Cl ₆ O |
| Amide | 180409-60-3 | Cyflufenamid | C ₂₀ H ₁₇ F ₅ N ₂ O ₂ | PCDF | 67562-39-4 | 1,2,3,4,6,7,8-HpCDF | C ₁₂ HCl ₇ O |
| Strobilurin | 117428-22-5 | Picoxystrobin | C ₁₈ H ₁₆ F ₃ NO ₄ | PCDF | 70648-26-9 | 1,2,3,4,7,8-HxCDF | C ₁₂ H ₂ Cl ₆ O |
| Strobilurin | 131860-33-8 | Azoxystrobin | C ₂₂ H ₁₇ N ₃ O ₅ | PCDF | 72918-21-9 | 1,2,3,7,8,9-HxCDF | C ₁₂ H ₂ Cl ₆ O |
| Strobilurin | 141517-21-7 | Trifloxystrobin | C ₂₀ H ₁₉ F ₃ N ₂ O 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 ₂ O ₃ | Phthalate | 84-66-2 | Diethyl | C ₁₂ H ₁₄ O ₄ |
| Strobilurin | 175013-18-0 | Pyraclostrobin | C ₁₉ H ₁₈ ClN ₃ O ₄ | Phthalate | 84-74-2 | Dibutyl | C ₁₆ H ₂₂ O ₄ |
| Strobilurin | 361377-29-9 | Fluoxastrobin | C ₂₁ H ₁₆ ClFN ₄ O ₅ | Phthalate | 85-68-7 | Benzylbutyl | C ₁₉ H ₂₀ O ₄ |
| Triazine | 122-34-9 | Simazine | C ₇ H ₁₂ ClN ₅ | Phthalate | 117-81-7 | Di-2-ethylhexyl | C ₂₄ H ₃₈ O ₄ |
| Triazine | 834-12-8 | Ametryn | C ₉ H ₁₇ N ₅ S | Phthalate | 117-84-0 | Di-n-octyl | C ₂₄ H ₃₈ O ₄ |
| Triazine | 886-50-0 | Terbutryn | C ₁₀ H ₁₉ N ₅ S | Phthalate | 131-11-3 | Dimethyl | C ₁₀ H ₁₀ O ₄ |
| Triazine | 1912-24-9 | Atrazine | C ₈ H ₁₄ ClN ₅ | Phthalate | 26761-40-0 | Di-isodecyl | C ₂₈ H ₄₆ O ₄ |
| Triazine | 5915-41-3 | Terbuthylazine | C ₉ H ₁₆ ClN ₅ | Phthalate | 28553-12-0 | Di-isononyl | C ₂₆ H ₄₂ O ₄ |
| Triazine | 21725-46-2 | Cyanazine | C ₆ H ₁₃ ClN ₆ | | | | |
| Triazine | 66215-27-8 | Cyromazine | C ₆ H ₁₀ N ₆ | PCDD | 1746-01-6 | 2,3,7,8-tetrachloro-dibenzo-p-dioxine | C ₁₂ H ₄ Cl ₄ O ₂ |
| Diazine | 1698-60-8 | Chloridazon | C ₁₀ H ₈ ClN ₃ O | PCDD | 3268-87-9 | OCDD | C ₁₂ Cl ₈ O ₂ |

| | | | | | | | |
|--------------------------|-------------|-----------------|---|----------|------------|-----------------------|---|
| Diazine | 25057-89-0 | Bentazone | C ₁₀ H ₁₂ N ₂ O ₃ S | PCDD | 19408-74-3 | 1,2,3,7,8,9-HxCDD | C ₁₂ H ₂ Cl ₆ O ₂ |
| Triazinone | 21087-64-9 | Metribuzin | C ₈ H ₁₄ N ₄ OS | PCDD | 35822-46-9 | 1,2,3,4,6,7,8-HpCDD | C ₁₂ HCl ₇ O ₂ |
| Triazinone | 41394-05-2 | Metamitron | C ₁₀ H ₁₀ N ₄ O | PCDD | 39227-28-6 | 1,2,3,4,7,8-HxCDD | C ₁₂ H ₂ Cl ₆ O ₂ |
| Triazole | 61-82-5 | Amitrole | C ₂ H ₄ N ₄ | PCDD | 40321-76-4 | 1,2,3,7,8-PeCDD | C ₁₂ H ₃ Cl ₅ O ₂ |
| Triazole | 76674-21-0 | Flutriafol | C ₁₆ H ₁₃ F ₂ N ₃ O | PCDD | 57653-85-7 | 1,2,3,6,7,8-HxCDD | C ₁₂ H ₂ Cl ₆ O ₂ |
| Triazole | 94361-06-5 | Cyproconazole | C ₁₅ H ₁₈ ClN ₃ O | | | | |
| Triazole | 119446-68-3 | Difenoconazole | C ₁₉ H ₁₇ Cl ₂ N ₃ O ₃ | Medicine | 298-46-4 | Carbamazepine | C ₁₅ H ₁₂ N ₂ O |
| Triazole | 131983-72-7 | Triticoconazole | C ₁₇ H ₂₀ ClN ₃ O | Medicine | 14168-01-5 | Dilor | C ₁₀ H ₇ Cl ₇ |
| Triazole | 133855-98-8 | Epoxiconazole | C ₁₇ H ₁₃ ClFN ₃ O | | | | |
| Thiocarbamate | 137-26-8 | Thiram | C ₆ H ₁₂ N ₂ S ₄ | Hormone | 50-28-2 | Estradiol | C ₁₈ H ₂₄ O ₂ |
| Thiocarbamate | 759-94-4 | EPTC | C ₉ H ₁₉ NOS | | | | |
| Thiocarbamate | 1929-77-7 | Vernolate | C ₁₀ H ₂₁ NOS | Auxin | 87-51-4 | Indolylacetic acid | C ₁₀ H ₉ NO ₂ |
| Thiocarbamate | 2303-16-4 | Di-allate | C ₁₀ H ₁₇ Cl ₂ NOS | | | | |
| Thiocarbamate | 2303-17-5 | Tri-allate | C ₁₀ H ₁₆ Cl ₃ NOS | Other | 608-73-1 | Hexachlorocyclohexane | C ₆ H ₆ Cl ₆ |
| Chloroacetamide | 1918-16-7 | Propachlor | C ₁₁ H ₁₄ ClNO | Other | 2550-75-6 | Chlorbicyclene | C ₉ H ₆ Cl 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 ₃ O | | | | |
| Dinitroaniline | 1582-09-8 | Trifluralin | C ₁₃ H ₁₆ F ₃ N ₃ O ₄ | | | | |
| Dinitroaniline | 19044-88-3 | Oryzalin | C ₁₂ H ₁₈ N ₄ O ₆ S | | | | |
| Dinitroaniline | 33629-47-9 | Butralin | C ₁₄ H ₂₁ N ₃ O ₄ | | | | |
| Dinitroaniline | 40487-42-1 | Pendimethalin | C ₁₃ H ₁₉ N ₃ O ₄ | | | | |
| Pyrethroid | 52315-07-8 | Cypermethrin | C ₂₂ H ₁₉ Cl ₂ NO ₃ | | | | |
| Pyrethroid | 52645-53-1 | Permethrin | C ₂₁ H ₂₀ Cl ₂ NO ₃ | | | | |
| Pyrethroid | 68359-37-5 | Cyfluthrin | C ₂₂ H ₁₈ Cl ₂ FNO ₃ | | | | |
| Triketone | 99105-77-8 | Sulcotrione | C ₁₄ H ₁₃ ClO ₅ S | | | | |
| Triketone | 335104-84-2 | Tembotrione | C ₁₇ H ₁₆ ClF ₃ O ₆ S | | | | |
| Phthalimide | 133-06-2 | Captan | C ₉ H ₈ Cl ₃ NO ₂ S | | | | |
| Phthalimide | 133-07-3 | Folpet | C ₉ H ₄ Cl ₃ NO ₂ S | | | | |
| Cyclodiene | 309-00-2 | Aldrine | C ₁₂ H ₈ Cl ₆ | | | | |
| Cyclodiene | 465-73-6 | Isodrine | C ₁₂ H ₈ Cl ₆ | | | | |
| Aryloxyalkanoic acid | 94-74-6 | MCPA | C ₉ H ₉ ClO ₃ | | | | |
| Aryloxyalkanoic acid | 7085-19-0 | Mecoprop | C ₁₀ H ₁₁ ClO ₃ | | | | |
| Alkylchlorophenoxy | 94-75-7 | 2,4-D | C ₈ H ₆ Cl ₂ O ₃ | | | | |
| Phosphonoglycine | 1071-83-6 | Glyphosate | C ₃ H ₇ NO ₃ P | | | | |
| Chloronitrile | 1897-45-6 | Chlorothalonil | C ₈ Cl ₄ N ₂ | | | | |
| Benzoic acid | 1918-00-9 | Dicamba | C ₈ H ₆ Cl ₂ O ₃ | | | | |
| Pyridine | 1918-02-1 | Picloram | C ₆ H ₃ Cl ₃ N ₂ O ₂ | | | | |
| Sulfite ester | 2312-35-8 | Propargite | C ₁₉ H ₂₆ O ₄ S | | | | |
| Ethylene generator | 16672-87-0 | Ethephon | C ₂ H ₆ ClO ₃ P | | | | |
| Dicarboximide | 36734-19-7 | Iprodione | C ₁₃ H ₁₃ Cl ₂ N ₃ O ₃ | | | | |
| Aryloxyphenoxypropionate | 51338-27-3 | Diclofop-methyl | C ₁₆ H ₁₄ Cl ₂ O ₄ | | | | |
| Diphenyl ether | 74070-46-5 | Aclonifen | C ₁₂ H ₉ ClN ₂ O ₃ | | | | |
| Anilinopyrimidine | 121552-61-2 | Cyprodinil | C ₁₄ H ₁₅ N ₃ | | | | |

| | | | |
|--------------------|--------------------|-------------|------------------------|
| Hydroxylanilide | 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 | $C_{21}H_{23}F_2NO_2$ |

1 **Table A2**

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

5
 6

| CAS number | Chemical formula | Parent CAS number | Parent name |
|--------------------|---|-----------------------|------------------------|
| 64-19-7 | CH ₃ COOH | 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 ₈ Cl ₄ | 50-29-3 | p,p'-DDT |
| 85-41-6 | C ₈ H ₅ NO ₂ | 133-07-3 | Folpet |
| 88-97-1 | C ₈ H ₇ NO ₃ | 133-07-3; 131983-72-7 | Folpet; Triticonazole |
| 88-99-3 | C ₆ H ₄ (COOH) ₂ | 133-07-3 | Folpet |
| 95-76-1 | C ₆ H ₅ Cl ₂ N | 330-54-1 | Diuron |
| 1024-57-3 | C ₁₀ H ₅ Cl ₇ O | 57-74-9; 76-44-8 | Chlordane; Heptachlore |
| 1031-07-8 | C ₉ H ₆ Cl ₆ O ₄ S | 115-29-7 | Endosulfan |
| 1570-64-5 | C ₇ H ₇ ClO | 94-74-6 | MCPA |
| 1897-45-6-INRA-1 | C ₈ HCl ₃ N ₂ O | 1897-45-6 | Chlorothalonil |
| 1897-45-6-INRA-2 | C ₈ H ₃ Cl ₃ N ₂ O ₄ S | 1897-45-6 | Chlorothalonil |
| 2327-02-8 | C ₇ H ₆ Cl ₂ N ₂ O | 330-54-1 | Diuron |
| 3567-62-2 | C ₈ H ₈ Cl ₂ N ₂ O | 330-54-1 | Diuron |
| 3739-38-6 | C ₁₃ H ₁₀ O ₃ | 52315-07-8 | Cypermethrin |
| 27304-13-8 | C ₁₀ H ₄ Cl ₈ O | 57-74-9 | Chlordane |
| 34256-82-1-INRA-1 | C ₁₄ H ₂₁ NO ₅ S | 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 ₅ S | 34256-82-1 | Acetochlor |
| 63637-89-8 | C ₁₇ H ₂₀ ClN ₃ O ₂ | 36734-19-7 | Iprodione |
| 68359-37-5-INRA-1 | C ₈ H ₁₀ Cl ₂ O ₂ | 68359-37-5 | Cyfluthrin |
| 77279-89-1 | C ₁₃ H ₉ FO ₃ | 68359-37-5 | Cyfluthrin |
| 131983-72-7-INRA-1 | C ₁₇ H ₂₀ ClN ₃ O ₂ | 131983-72-7 | Triticonazole |
| 131983-72-7-INRA-2 | C ₁₇ H ₂₀ ClN ₃ O ₂ | 131983-72-7 | Triticonazole |

7
 8
 9
 10
 11
 12

1 **Table A3**

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

6

| 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 | Etoazole |
| 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 |

1
2