Clustering pesticides according to their molecular properties, fate, and effects by considering additional ecotoxicological parameters in the TyPol method

Harouna Traoré, Olivier Crouzet, Laure Mamy, Christine Sireyjol, Virginie Rossard, Rémi Servien, Eric Latrille, Fabrice Martin-Laurent, Dominique Patureau, Pierre Benoit

Clustering pesticides according to their molecular properties, fate and effects by considering additional ecotoxicological parameters in the TyPol method

Harouna Traoré[1, 2] · Olivier Crouzet[2] · Laure Mamy[1] · Christine Sireyjol[2] · Virginie Rossard[3] · Rémi Servien[4] · Eric Latrille[3] · Fabrice Martin-Laurent[5] · Dominique Patureau[3] · Pierre Benoit[1]

[1] UMR ECOSYS, INRA, AgroParisTech, Université Paris-Saclay, 78850 Thiverval-Grignon, France

[2] UMR ECOSYS, INRA, AgroParisTech, Université Paris-Saclay, 78206 Versailles, France

[3] UR LBE, INRA, Université Montpellier11100 Narbonne, France

[4] INRA, UMR Toxalim, 31300 Toulouse, France

[5] INRA, UMR AgroEcologie 21065 Dijon, France

✉ Pierre Benoit

pierre.benoit@inra.fr

Traoré H, Crouzet O, Mamy L, Sireyjol C, Rossard V, Servien R, Latrille E, Benoit P, 2018. Clustering pesticides according to their molecular properties and their fate and effects by considering additionnal ecotoxicological parameters in the TyPol method. **Environmental Science and Pollution Research, 25: 4728-4738.**

https://doi.org/10.1007/s11356-017-0758-8

Abstract

Understanding the fate and ecotoxicological effects of pesticides largely depends on their molecular properties. We recently developed TyPol (Typology of Pollutants), a classification method of organic compounds based on statistical analyses. It combines several environmental (sorption coefficient, degradation half-life) and one ecotoxicological (bioconcentration factor) parameters, to structural molecular descriptors (number of atoms in the molecule, molecular surface, dipole moment, energy of orbitals, etc.). The present study attempts to extend TyPol to the ecotoxicological effects of pesticides on non-target organisms, based on data analysis from available literature and databases. It revealed that relevant ecotoxicological endpoints for terrestrial organisms (e.g., soil microorganisms, invertebrates) that support a range of ecosystemic services are lacking as compared to aquatic organisms. The availability of ecotoxicological parameters was also lower for chronic than for acute ecotoxicity endpoints. Consequently, seven parameters were included for acute (EC50, LC50) and chronic (NOEC) ecotoxicological effects for one terrestrial (*Eisenia* sp.) and three aquatic (*Daphnia* sp., algae, *Lemna* sp.) organisms. In this new configuration, we used TyPol to classify 50 pesticides into different clusters that gather molecules with similar environmental behaviors and ecotoxicological effects. The classification results evidenced relationships between molecular descriptors, environmental parameters, and the added ecotoxicological endpoints. This proof-of concept study also showed that TyPol in silico classification can successfully address new scientific questions and be expanded with other parameters of interest.

Keywords Pesticides · Molecular descriptor · Fate · Ecotoxicity · Clustering

Introduction

Assessing and linking the fate and ecotoxicological effects of organic contaminants in the environment remains a major time- and resource-intensive challenge. Therefore, advances in our understanding of the (eco)toxicological impacts of the thousands of chemicals being produced or already marketed are often painfully slow and lag significantly behind the public's demand for information. Reliable measurements of environmental fates and effects, and risk assessment procedures, strongly rely on the ability to accurately measure or estimate various environmental parameters and ecotoxicological endpoints (Sabljic 2001). According to the literature, 30,000 to 100,000 chemical substances may be concerned by environmental risk assessment (ERA) (Muir and Howard 2006). However, they cannot be studied on a case-by-case basis, in particular because experimental studies are time-consuming and/or cost-prohibitive (Russom et al. 2003). In such a context, the development of in silico methods of prediction based on Quantitative Structure Activity Relationships (QSARs) (or Quantitative Structure Property Relationships, QSPRs) has received increasing attention for many years (Cronin et al. 2003; Hermens et al. 1995; Lapenna et al. 2010).

QSARs estimate one or several properties of compounds (such as sorption on soils and sediments, biodegradation, bioconcentration factor, or biological activities) from other properties such as structural molecular properties (molecular weight, molecular surface, energies, etc.), water solubility, or the octanol-water partition coefficient (e.g., OECD 1993; Worrall 2001; Sabljic and Nakagawa 2014; Mamy et al. 2015). An important number of QSARs have recently been proposed for predicting behavioral or toxicological parameters (Eriksson et al. 2002; Pavan et al. 2008; Lapenna et al. 2010; Mamy et al. 2015). Other approaches aim at ranking organic compounds based on the values of several of their properties such as partitioning, persistence, or bioaccumulation. Therefore, approaches able to classify compounds according to their environmental behavior and/or (eco)toxicological effects will help regulators and scientists to tackle the constant increase in the diversity and number of chemical substances.

We recently developed "TyPol" (Typology of Pollutants), a classification method based on statistical analyses combining several environmental parameters (i.e., sorption coefficient, degradation half-life, Henry constant) and an ecotoxicological parameter (bioconcentration factor BCF), and structural molecular descriptors (i.e., number of atoms in the molecule, molecular surface, dipole moment, energy of orbitals). Molecular descriptors are calculated using an in silico approach. Environmental parameters and BCF are extracted from available databases and literature (Servien et al. 2014). So far, TyPol has been mainly focused on pesticides and their transformation products. It can infer possible changes in environmental fate following different degradation

processes through modifying the clustering as compared to the parent compounds (Servien et al. 2014; Benoit et al. 2017). As indicated above, the initial version of TyPol did not include any ecotoxicological endpoints except the bioconcentration factor, which informs more on the transfer along the food chain than on the effects on non-target organisms.

The objective of this study was to implement the TyPol pesticide database with a set of ecotoxicological data for several aquatic and terrestrial organisms. We first carried out a data analysis based on available literature and pesticide databases, and focused on ecotoxicological endpoints recommended by regulatory risk assessment guidelines for terrestrial and aquatic organisms (microorganisms, plants, and invertebrates). The review resulted in the selection of seven parameters for acute (EC50, LC50) and chronic (NOEC) effects of 50 pesticides on earthworms (*Eisenia*), crustaceans (*Daphnia* sp.), algae, and lemnoideae (*Lemna* sp.). These parameters were implemented in the database, and in this new configuration, TyPol was used to classify pesticides into different clusters as a proof of concept. Finally, classification results were analyzed to reveal relationships between molecular descriptors, environmental parameters, and the added ecotoxicological parameters.

Material and methods

Selection of ecotoxicological parameters

We first considered the 146 pesticide-related molecules (122 active substances and 24 transformation products (TPs)) already present in the TyPol database for which environmental parameters, BCF, and molecular descriptors had previously been implemented (Servien et al. 2014). To include the ecotoxicological effects of pesticides on non-target organisms in TyPol, we selected biological models with available ecotoxicological endpoint data. We primarily decided to focus on terrestrial/aquatic invertebrates and aquatic plants because standardized bioassays are included in regulatory guidelines for pesticide registration (EU-Regulation 1107/2009/EC) or risk assessment (Table 1). Thus, we first used databases such as PPDB and Agritox (Agritox 2017; PPDB 2017). In addition, we consulted databases such as INERIS (2017), and the US-EPA ECOTOX Knowledgebase (2017), and also regulatory reports about active substances from the ECHA and EFSA European Agencies, in which validated ecotoxicological endpoints are available. We performed a complementary literature survey to collect missing values. Acute and chronic effects were both considered, and only the well-documented molecules were retained. One difficulty was to find data obtained for a same species and according to the same protocols. Especially for algal tests, different *Chlorophyceae* species (*Raphidocelis subcapitata*, *Scenedesmus subspicatus*) are used according to ISO standard recommendations (Table 1), so we decided to gather data concerning *Raphidocelis*

subcapitata and *Scenedesmus subspicatus*. For earthworms, we found LC50 and NOEC values from either *Eisenia fetida* or *Eisenia andrei*. Microbial ecotoxicological data were collected from Microtox® (ISO 11348 2009) and from the standardized nitrification assay (ISO 15685 2012). Although not included in pesticide ERA, these tests are frequently used in microbial ecotoxicology. However, for these two tests, too many values were missing to be included in our study.

Table 1. Selected non-target organisms and corresponding acute and chronic ecotoxicological parameters

Organisms	Normalized bio-assay	Ecotoxicological parameters
Earthworms	*Eisenia fetida*	LC50 acute (µmol/kg), 14 or 28 d – survival (ISO 11268-1, OCDE 207)
	Eisenia andrei	NOEC chronic (µmol/kg), 28 d – reproduction (ISO 11268-2, OCDE 222)
Aquatic invertebrates	*Daphnia* sp.	EC50 acute (µmol/L), 48 h – mobility test (ISO 6341, OCDE 202)
		NOEC chronic (µmol/L), 21 d – mobility test (ISO 10706, OCDE 211)
Algae	*Pseudokirchneriella subcapitata (*or *Selenastrum capricornus* or *Raphidocelis subcapitata)*	EC50 acute (µmol/L), 72 or 96 h – growth or biomass (ISO 8692, OCDE 201)
		NOEC chronic (µmol/L), 72 or 96 h – growth or biomass (ISO 8692, OCDE 201)
Aquatic plants	*Lemna* sp.	EC50 acute (µmol/L), 7 d – biomass (ISO 20079, OCDE 221)

TyPol already included the bioconcentration factor (BCF), which describes bioaccumulation across trophic networks (Servien et al. 2014). We implemented seven new ecotoxicological parameters to cover both acute and chronic effects of pesticides on non-target soil and aquatic organisms. For acute toxicity, we used LC50 for earthworms (*Eisenia* sp.), and EC50 for aquatic invertebrates in surface waters (*Daphnia* sp.), aquatic plants (*Lemna* sp.) and unicellular algae (*Chlorophyceae* sp.). Chronic effects were described by NOEC for *Eisenia* sp., *Daphnia* sp., and algae.

Environmental parameters

Six environmental parameters are addressed in TyPol: water solubility (Sw) and the octanol-water partition coefficient (K_{ow}) to describe dissolution; vapor pressure (P_{vap}) for volatilization from soil and plants; Henry's law

constant (K_H) for volatilization from water; the adsorption coefficient normalized to soil carbon organic content (K_{oc}) for adsorption, and half-life (DT50) for degradation (Servien et al. 2014).

Clustering methodology

The TyPol tool classifies pesticides and their transformation products according to their behavior in the environment and their molecular properties (Servien et al. 2014; Benoit et al. 2017). The strategy relies on partial least squares (PLS) analysis and hierarchical clustering (see "Statistical treatments" section below). We assessed its robustness on a list of 215 organic compounds using a cross-validation algorithm. The information system is based on MySQL DBMS-R management system for relational database (version 5.1), the Apache web server (version 2.2), and the statistical R software program (also used for graphs). More details concerning TyPol can be found in Servien et al. (2014).

Statistical treatments

In TyPol, the PLS model is applied to find the multidimensional directions X in the observable variable (molecular descriptor) space that explains the maximum multidimensional variance Y direction in the predicted variable (environmental and ecotoxicological parameter) space. The optimal number p of PLS components to perform clustering is selected according to Wold rules (Wold 1978). We also used the NIPALS (Non-linear Iterative PArtial Least Squares) algorithm, which performs PLS without removing individuals with missing values and without estimating these missing values (Tenenhaus 1998). Nevertheless, reducing the number of missing values as much as possible is recommended to obtain the most accurate results. After PLS analysis, we carried out a hierarchical clustering algorithm on the PLS X1,…, Xp and Y1,…, Yp axes to categorize the molecules by assigning similar compounds to one cluster. At each step, the final number of clusters was chosen by comparing heights in the dendrogram, a statistical map summarizing Ward clustering. Minimization of intra-variability and maximization of intervariability were retained to choose the most appropriate number of clusters (Servien et al. 2014).

Results and Discussion

Available data for ecotoxicological effects on soil and water systems

Contrarily to our expectations, although 122 active substances and almost all of the 24 transformation products (TPs) preexisting in the TyPol database were indexed in the PPDB and Agritox databases, we failed to retrieve the selected ecotoxicological endpoints for all of them (Supplementary material - Table A1). We found the seven

selected parameters in the PPDB or Agritox databases only for chlorothalonil, epoxiconazole, glyphosate, metribuzin, pendimethalin, tebuconazole, triallate, and trifluralin, and for only one TP, toxanilic acid (derived from metolachlor). Conversely, data were completely missing for the following active substances: di-allate, fosetyl, kelevan, metsulfuron, and neburon, and also for several TPs: 1-(3,4-dichlorophenyl)-3-methylurea, 3-ketocarbofuran, 4-fluoro-3-phenoxybenzoic acid, and desmethyl norflurazon. In general, TP data were much less available in both databases. But even if we excluded TPs and only considered the 122 active substances, 5 or 6 ecotoxicological endpoints were still missing for a significant number of active substances (Supplementary material Table A1).

We found strong variability in data availability according to ecotoxicological parameters and biological models. Acute toxicity endpoints were the most available data for aquatic invertebrates (*Daphnia* sp.), with values present in at least one of the databases for 114 active substances. By contrast, chronic effects on terrestrial invertebrates (*Eisenia* sp.) were the least available data, with values found only for 32 active substances. The same trends were observed for TPs. Again the least abundant data concerned chronic toxicity effects, with missing values for 88, 96, and 92% of the transformation products for *Daphnia* sp., algae, and earthworms, respectively. For acute toxicity, EC50 values were much more available for *Daphnia* sp. (22% of missing values for the selected TPs) than for algae, *Lemna* sp., or earthworms (68, 84, and 68% of missing values, respectively).

In order to reduce the number of missing values and to perform TyPol analysis, we selected a set of 50 pesticides with the highest number of available data for the next step. Then we conducted a detailed survey of the selected pesticides in other databases and literature reviews (US-EPA ECOTOX, ECHA and EFSA reports, Web of Science) to gather the remaining missing data (Table 2; see the detailed reference list in Supplementary Material). Thanks to this data mining, we retrieved almost all ecotoxicological endpoints for the 50 molecules, and the level of missing data was reduced to 3.4% of missing values for the whole dataset (Table 2).

Table 2. Ecotoxicological parameters obtained from normalized bioassays and implemented in TyPol database for the 50 selected pesticides (AS: active substance, PGR: Plant growth regulator)

Chemical family	CAS number	Name	Pesticide type	*Daphnia* sp.		*Lemna* sp.	*P. subcapitata (*or other *chlorphyceae)*		*Eisenia fetida* *Eisenia andrei*	
				EC50	NOEC	EC50	EC50	NOEC	LC50	NOEC
				µmol $_{(AS)}$ L^{-1}					µmol $_{(AS)}$ kg^{-1}	
Amide	15299-99-7	Napropamide	Herbicide	52,7 a	15,8 a	0,884 a	12,5 a	4,42 b	1039 a	111 a
Amide	118134-30-8	Spiroxamine	Fungicide	10,1 c	0,114 c	9,28 a	0,011 a	0,007 a	1681 a	108 a
Arylophenoxy acid	94-75-7	2,4-D	Herbicide, PGR	452 d	173,7 d	2,62 d	109 a	86,9 d	1583 a	283 d
Benzoic acid	1918-00-9	Dicamba	Herbicide	403 e	439 a	53,8 e	45,2 e	16,7 e	4524 a	Not found
Carbamate	16118-49-3	Carbetamide	Herbicide	343 a	4,23 f	1274 a	669 a	212 f	2793 a	Not found
Carbamate	1563-66-2	Carbofuran	Insecticide	0,172 b	0,044 b	1067 b	29,4 a	14,5 a	38,2 h	2,26 h
Carbamate	13684-56-5	Desmedipham	Herbicide	1,50 a	0,033 a	17,3 a	0,033 a	0,033 a	266 a	8,32 a
Carbamate	13684-63-4	Phenmedipham	Herbicide	1,37 a	0,107 a	0,766 a	0,633 i	0,083 i	147 j	7,21 j
Chloroacetamide	34256-82-1	Acetochlor	Herbicide	31,9 a	0,082 a	0,010 a	0,007 k	0,0002 k	391 a	74,1 l
Chloroacetamide	15972-60-8	Alachlor	Herbicide	37,1 a	0,853 a	0,037 a	0,004 m	0,001 m	1435 m	Not found
Chloroacetamide	51218-45-2	Metolachlor	Herbicide	82,8 a	2,50 a	0,152 a	0,195 n	0,134 n	2008 o	8,95 o
Chloroacetamide	67129-08-2	Metazachlor	Herbicide	80,3 p	0,360 a	0,008 a	0,112 q	0,007 a	1584 p	Not found
Chloronitrile	1897-45-6	Chlorothalonil	Fungicide	0,203 a	0,034 a	1,09 a	0,790 a	0,124 a	357 r	37,6 r
Cinamic acid	110488-70-5	Dimethomorph	Fungicide	27,3 s	1,19 s	56,8 s	107 s	25,3 s	2578 a	309 a
Diazine	25057-89-0	Bentazone	Fungicide	260 a	499 a	22,5 a	18,7 b	3,66 b	3620 a	Not found
Diazine	1698-60-8	Chloridazone	Herbicide	596 a	28.1 a	20.8 b	13.5 a	3.29 a	4513 a	Not found
Dicarboximide	36734-19-7	Iprodione	Fungicide	2,00 a	0,515 a	3,03 a	5,45 a	9,69 a	3029 a	Not found
Dinitroaniline	33629-47-9	Butralin	PGR, Herbicide	0,406 a	0,267 a	0,508 a	0,406 a	Not found	3386 a	84,3 a
Dinitroaniline	40487-42-1	Pendimethalin	Herbicide	0,523 a	0,052 a	0,078 a	0,021 a	0,011 a	3555 t	119 t
Dinitroaniline	1582-09-8	Trifluralin	Herbicide	0,731 u	0,152 a	0,130 a	0,036 a	0,015 a	1491 a	42,3 a
Diphenyl ether	74070-46-5	Aclonifen	Herbicide	4,53 a	0,060 a	0,023 a	0,106 v	0,026 v	567 a	170 a
Hydroxyanilide	126833-17-8	Fenhexamid	Fungicide	62,2 a	3,31 a	7,61 b	86,0 a	17,7 a	1655 a	33,1 a

Chemical family	CAS number	Name	Pesticide type	*Daphnia* sp.		*Lemna* sp.	*P. subcapitata (*or other *chlorphyceae)*		*Eisenia fetida* *Eisenia andrei*	
				EC50	NOEC	EC50	EC50	NOEC	LC50	NOEC
Neonicotinoid	135410-20-7	Acetamiprid	Insecticide	224 a	22,5 a	4,49 a	450 a	441	40 a	5,66 a
Organochlorine	115-29-7	Endosulfan	Insecticide	0,451 w	0,035 w	246 w	5,28 w	1,38 w	27 w	0,25 w
Organochlorine	58-89-9	Lindane	Insecticide	5,50 a	0,186 a	0,093 a	8,60 a	6,53 a	395 x, y	61,9 y
Organophosphorous	13194-48-4	Ethoprophos	Insecticide	0,825 a	0,008 a	413 a	115,6 a	13,2 a	163 a	34,3 a
Organophosphorous	122-14-5	Fenitrothion	Insecticide	0,031 a	0,0001 a	Not found	4,69 a	0,361 a	833 a	90,2 a
Phosphonoglycine	1071-83-6	Glyphosate	Herbicide	237 a	177,4 a	71,0 a	26,0 a	3,78 z	5914 z	2791 aa
Phthalimide	133-06-2	Captan	Fungicide	23,6 a	1,86 a	42,2 a	3,93 a	33,3 a	1726 a	40,6 a
Pyridine	01/02/1918	Picloram	Herbicide	142 a	48,9 a	422 a	157,4 a	33,1 b	20707 a	Not found
Strobilurin	131860-33-8	Azoxystrobin	Fungicide	0,570 a	0,109 a	7,93 a	0,892 a	0,094 ab	702 a	49,6 a
Strobilurin	361377-29-9	Fluoxastrobin	Fungicide	1.05 a	0.39 a	13.1 a	0.57 a	0.17 g	1090 a	2.90 a
Strobilurin	175013-18-0	Pyraclostrobin	Fungicide	0,04 a	0,01 a	4.44 a	0.39 a	0,04 ac	1462 a	1.14 a
Strobilurin	141517-21-7	Trifloxystrobin	Fungicide	0,027 a	0,007 a	4.73 a	0,09 a	0,02 a	982 ad	8,57 a
Thiocarbamate	2303-17-5	Tri-allate	Herbicide	1,41 a	0,043 a	7,55 a	0,007 a	0,105 a	901 a	44,7 a
s-Triazine	1912-24-9	Atrazine	Herbicide	72,3 ae	0,649 ae	0,394 af	0,621 af	0,070 af	1048 ag	46,4 ah
s-Triazine	66215-27-8	Cyromazine	Insecticide	602 a	27,7 a	Not found	746 a	602 a	6018 a	2004 a
s-Triazine	122-34-9	Simazine	Herbicide	5,45 a	12,4 a	0,69 b	0,496 b	0,159 b	4959 a	Not found
Triazinone	41394-05-2	Metamitron	Herbicide	480 ai	28,2 ai	2,0 a	0,692 aj	0,495 aj	4520 a	138 a
Triazinone	21087-64-9	Metribuzin	Herbicide	229 a	1,49 a	0,173 ak	0,201 ak	0,089 ak	1993 a	24,5 a
Triazole	94361-06-5	Cyproconazole	Fungicide	89,1 a	0,994 a	0,202 a	0,343 a	0,069 a	576 a	2,57 a
Triazole	133855-98-8	Epoxiconazole	Fungicide	26,4 a	1,91 a	0,042 a	3,61 a	0,237 a	1083 ad	0,51 a
Triazole	76674-21-0	Flutriafol	Fungicide	222 a	1,03 a	2,2 a	6,31 a	2,16 a	1660 a	Not found
Triazole	107534-96-3	Tebuconazole	Fungicide, PGR	9,06 a	0,032 a	0,8 am	12,3 am	0,325 am	4486 a	32,5 a
Triazole	131983-72-7	Triticonazole	Fungicide	28,3 a	0,29 a	4,4 a	3,15 an	7,87 b	1573 an	787 an
Triketone	99105-77-8	Sulcotrione	Herbicide	304 ao	228 ao	1,703 ao	10,646 ao	0,578 ao	3042 a	Not found
Triketone	335104-84-2	Tembotrione	Herbicide	111 ap	11,3 a	0,014 a	0,862 a	0,454 ap	2268 a	2,84 a
Substituted Urea	64902-72-3	Chlorsulfuron	Herbicide	1037 a	33,5 a	0,001 a	0,190 a	0,026 b	2795 a	524 a
Substituted Urea	330-54-1	Diuron	Herbicide	6,01 aq	0,429 aq	0,079 a	0,012 a	0,008 aq	3424 a	61,8 a

Chemical family	CAS number	Name	Pesticide type	*Daphnia* sp.		*Lemna* sp.	*P. subcapitata (*or other *chlorphyceae)*		*Eisenia fetida* *Eisenia andrei*	
				EC50	NOEC	EC50	EC50	NOEC	LC50	NOEC
Substituted Urea	34123-59-6	Isoproturon	Herbicide	2,81 a	0,582 a	0,150 a	0,155 ar	0,087 ar	4848 a	Not found

(a) PPDB 2017; (b) US-EPA ECOTOX Knowledgebase 2017; (c) ECHA report for Spiroxamine 2015; (d) EFSA 2014; (e) EFSA 2011; (f) ECHA 2014; (h) De Silva et al. 2009; (i) EC 1999; (j) PPDB and Van Gestel et al. 1992; EFSA 2011; (l) Xiao et al. 2006; (m) EC 2007; (n) PAN Pesticides Database 2017; (o) EU - Pesticides database 2017; (p) FAO 1999; (q) ECHA 2011; (r) Leitao et al. 2014; (s) EFSA 2004; (t) Belden et al. 2005; (u) EC 2005; (v) EC 2006; (w) Rotterdam Convention 2011; (x) PPDB, Haque et al. 1983, and Lock et al. 2002; (y) Lock et al. 2002; (z) EU 2002; (aa) von Mérey et al. 2016; (ab) EFSA 2009; (ac) US-EPA 2005; (ad)Wang et al. 2012; (ae) Giddings 2002; (af) UE 2004; (ag)Wang et al. 2012, and Haque and Ebing 1983; (ah) Song et al. 2009; (ai) Nitschke et al. 1999; (aj) Vaittinen 1987; (ak) Fairchild et al. 1997; (al) ECHA 2012; (am) EFSA 2005; (an) ECHA 2011; (ao) ECHA 2012; (ap) UE 2005; (aq) ECHA 2015. Details about these references are specified in supplementary materi

The fact that data about chronic effects of pesticides were less available is explained by current rules in EU pesticide registration procedures. As a matter of fact, results of standardized chronic toxicity tests are mandatory in case of continuous or repeated exposure, otherwise acute toxicity tests are sufficient. We should point out that for ecotoxicological assessment of soil organisms and aquatic plants, even if standardized tests do exist on model organisms, data on pesticide ecotoxicological effects are still dramatically missing. This statement is in line with recent scientific opinions on risk assessment of plant protection products for soil organisms and biodiversity (EFSA 2016a, 2017). It suggests implementing new tests on non-target organisms that act as keystones of ecosystem functions to fulfill ERA with specific goals as far as protecting ecosystem services is concerned. Therefore, ecotoxicological endpoints related to the behavior (Sousa et al. 2008), functions (i.e., EFSA 2016a), and recovery capabilities of non-target organism communities (EFSA 2016b) would deserve greater attention for more careful pesticide ERA.

Clustering with ecotoxicological parameters

Adding the seven new ecotoxicological parameters and considering optimal PLS conditions for only 50 pesticides prompted us to choose a number of PLS components equal to three according to Wold rules (Wold 1978). Figure 1 shows the projection of the data from the 50 pesticides on the first component axes. On this projection, clusters 2 and 3 strongly overlap, but they are better separated on the other component axes (data not shown). The composition of the four clusters is given in Table 3. The first component explained 35% of the variance of the molecular descriptors, and 34% of the variance of the environmental and ecotoxicological parameters. Its main characteristics were strong positive loadings for polarizability, geometric and topological descriptors, and constitutional descriptors like the number of carbon atoms and molecular mass (Fig. 2). In contrast, total energy had strong negative effects. The second axis explained only 6% of the variance of the molecular descriptors, and 32% of the variance of the environmental and ecotoxicological parameters. On this axis, variables such as the number of chlorine or halogen atoms had a positive loading, whereas LUMO energy, the number of multiple and aromatic bonds, or the number of oxygen atoms had a negative loading (Fig. 2).

We selected the number of clusters by plotting the heights of the dendrogram nodes and looking for a break. The best choice, which minimized intra-variability and maximized inter-variability, was to classify the compounds into four clusters. The largest cluster (cluster 4) contained 16 molecules (32% of the total), and the smallest (cluster 3) contained 8 molecules (16% of the total) (Table 3).

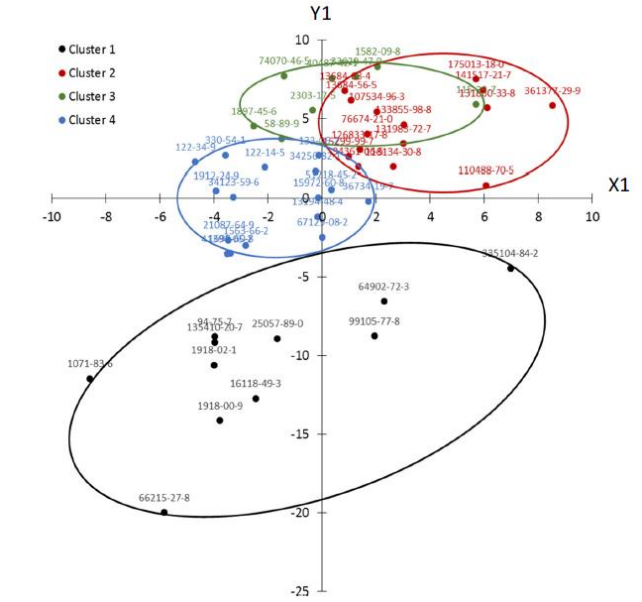

Fig. 1 Graphical representation Y1 of the four obtained clusters in the first component X, Y axes of the PLS with the molecular descriptors, and environmental and ecotoxicological parameters on the 50 selected pesticides

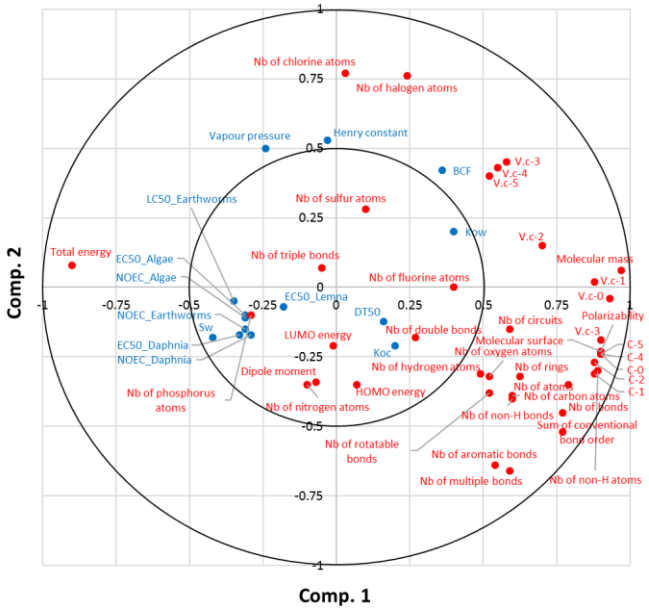

Fig. 2 Circles of correlations of the ecotoxicological and environmental parameters (in blue) and molecular descriptors (in red) variables on the two main components of the PLS. C.i–i stands for the connectivity index C.i of order i (i = 0 to 5), and V.c-i stands for the valence connectivity index V.c of order i (i = 0–5)

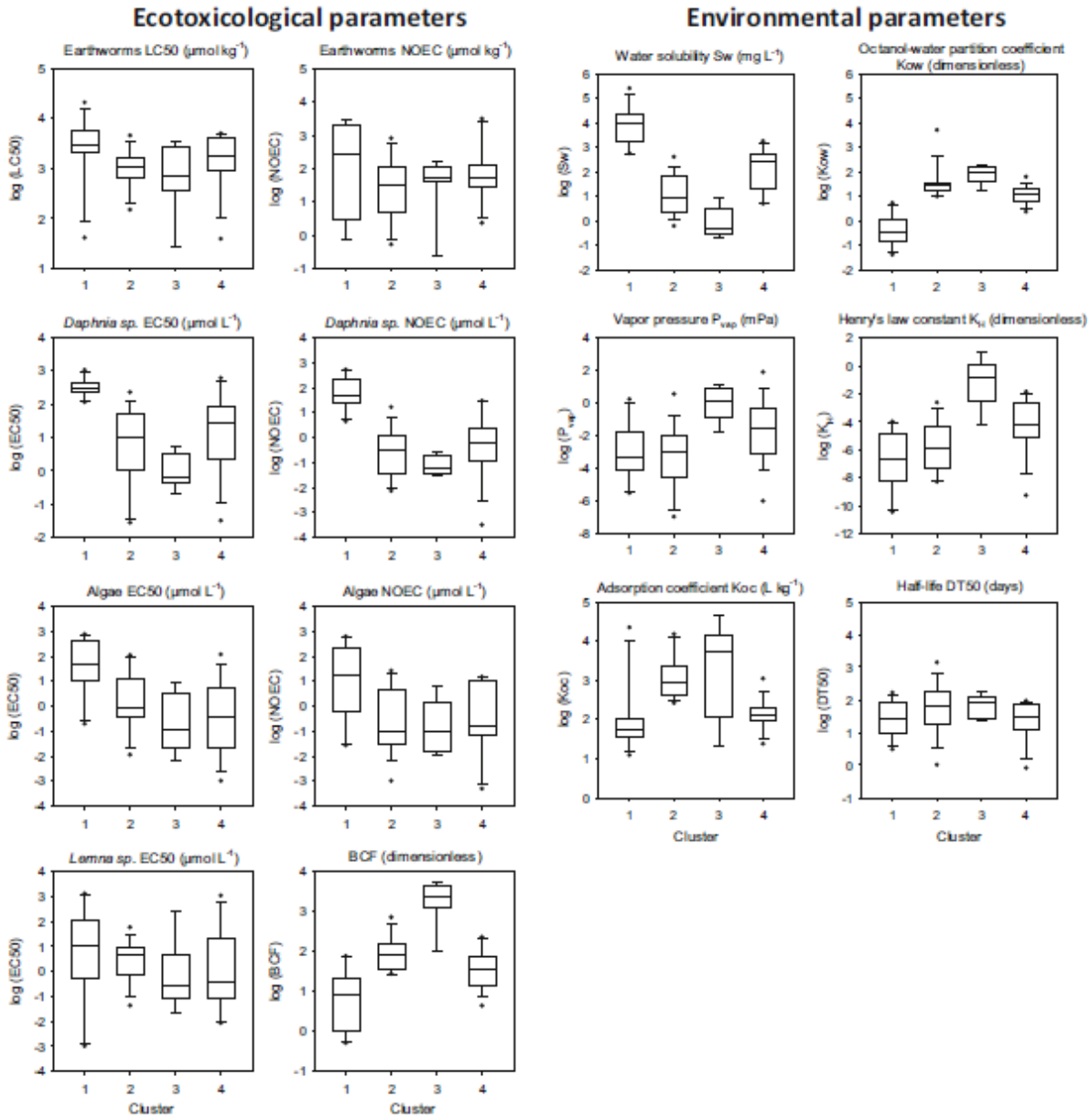

Fig. 3 Ranges of variation (box-and-whisker plots) of the values of ecotoxicological and environmental parameters considered in TyPol for the 50 selected pesticides classified in the four clusters: cluster 1 (11 pesticides), cluster 2 (15 pesticides), cluster 3 (8 pesticides) and cluster 4 (16 pesticides)

Table 3. TyPol clustering of the 50 selected pesticides using the molecular descriptors, and environmental and ecotoxicological parameters

Composition of each cluster			
Cluster 1	Cluster 2	Cluster 3	Cluster 4
Glyphosate	Tebuconazole	Endosulfan	Fenitrothion
Acetamiprid	Dimethomorph	Trifluralin	Simazine
Carbetamid	Spiroxamine	Chlorothalonil	Ethoprophos
Dicamba	Fenhexamid	Tri-allate	Captan
Picloram	Azoxystrobin	Butraline	Carbofuran
Bentazone	Triticonazole	Pendimethalin	Alachlore
Tembotrione	Epoxiconazole	Lindane	Chloridazon
Chlorsulfuron	Desmedipham	Aclonifen	Atrazine
Cyromazine	Phenmedipham		Metribuzine
2,4-D	Trifloxystrobin		Diuron
Sulcotrione	Napropamide		Isoproturon
	Pyraclostrobin		Acetochlor
	Fluoxastrobin		Metamitron
	Flutriafol		Metolachlor
	Cyproconazole		Metazachlor
			Iprodione

Identification of discriminating parameters and molecular descriptors

Variation ranges within each cluster are shown in Fig. 3 for all the ecotoxicological/environmental parameters and in Fig. 4 for selected molecular descriptors. Median values are given in Table A2 (Supplementary Material). Figure 3 clearly shows that the four clusters had close DT50 values and that their ranges overlapped to a high degree. It suggests that this parameter poorly discriminated the different molecules within the classification.

Cluster 1 gathered together phenoxyacids, glyphosate, and other polar pesticides (Table 3). This cluster included compounds characterized by a strong dipole moment and high total energy values, whereas polarizabilities (Fig. 4) and the different connectivity indices were low (Table A2). These compounds had low molecular weights and low Connolly molecular surface areas (Fig. 4). Concerning the different ecotoxicological parameters, this cluster was characterized by the lowest BCF and the highest NOEC and EC50 values (Fig. 3). This suggests that the compounds gathered in cluster 1 are lowly likely to bioaccumulate in living organisms (Regulation EC 1107/2009 2009) and also have low acute toxicity on invertebrates (median LC50 in earthworms = 3042 µmol/kg; median EC50 in *Daphnia* sp. = 304.2 µmol/L) or aquatic microorganisms (median EC50 in algae

= 45.2 µmol/L). The presence of glyphosate in this low toxicity cluster can appear as surprising because recent publications showed that glyphosate isopropylamine salt (CAS 38641-94-0) may be potentially harmful to the environment and biota (Cuhra et al. 2013). This discrepancy is due to the use of fairly high values of EC50 and NOEC recorded only for the active ingredient glyphosate (CAS 1071-83-6) in EU registered data compiled in the PPDB database and the other two references (Table 2). The highest median value was Sw (10,500 mg/L), while K_{oc} was the lowest (55.9 L/kg). These pesticides had low DT50 values (Fig. 3). This suggests that they have low persistence in the environment, and partly explains their low potential for bioaccumulation and toxicity (Regulation EC 1107/2009 2009). Finally, the compounds gathered in cluster 1 had low K_H values (i.e., low volatility from water) and low P_{vap} values, hence low volatility from soil and plants, and a low potential of transfer to the atmosphere (FOCUS 2008).

Cluster 2 mostly contained azoles, amides (fenhexamid, napropamide) and strobilurin fungicides, and a few carbamate (desmedipham, phenmedipham) herbicides (Table 3). These compounds had the lowest total energy values and the highest polarizability values, numbers of atoms, and connectivity indices (Fig. 4 and Table A2). Median Sw values were low (9.3 mg/L), which is consistent with high median K_{oc} values (885 L/kg) indicating a relatively high sorption potential. These molecules also had low K_H and P_{vap} values. Ecotoxicological endpoints indicated higher toxicity for the different non-target organisms than in cluster 1. Medium to high BCF values indicated their potential to significantly bioaccumulate in living organisms. Chronic toxicity was relatively high for algae (median NOEC value = 0.09 µmol/L), *Daphnia* sp. (median NOEC value = 0.29 µmol/L), and earthworms (median NOEC value = 20.5 µmol/kg) as compared to the other groups. Acute toxicity was also in the medium to high range considering the relatively low EC50 and LC50 values (Fig. 3).

Cluster 3 was composed of pesticides belonging to various chemical families (e.g., organochlorine, dinitroaniline, chloronitrile) and various categories of uses (insecticides, fungicides, and herbicides) (Table 3). It contained the most persistent ones, with a median DT50 value of 86 days. These molecules had low total energy values and relatively high polarizability values, numbers of atoms, Connolly molecular surface areas, and connectivity indices (Fig. 4). They had the lowest dipole moment values and the highest numbers of chlorine atoms (2), which is in line with the highest ecotoxicity, whether acute or chronic. EC50 median values were 0.22, 0.32, and 0.63 µmol/L for algae, *Lemna* sp. and *Daphnia* sp., respectively. NOEC values were the lowest for algae and *Daphnia* sp., whereas this chronic toxicity parameter was relatively high for earthworms as compared to cluster 2 (Fig. 3). BCF values were much higher than in the other clusters. In relation to this strong bioaccumulation potential, these compounds had the lowest water solubility (median Sw = 0.57 mg/L) and the highest octanol water

coefficient (median K_{ow} = 97.3). Accordingly, sorption coefficient values were high (median K_{oc} = 5713 L/kg). Their volatilization potential from water (K_H) and from soil or plants (P_{vap}) was also the highest, much higher than for compounds from the other clusters.

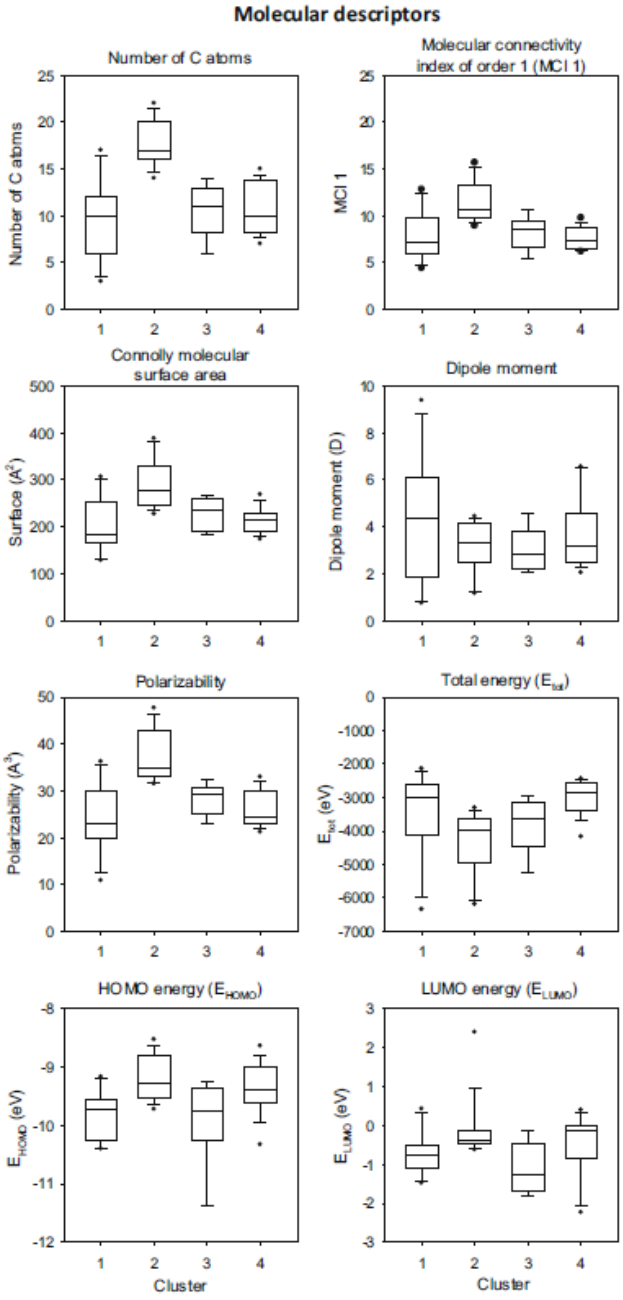

Fig. 4 Ranges of variation (box-and-whisker plots) of the values of selected molecular descriptors considered in TyPol for the 50 selected pesticides classified in the four clusters: cluster 1 (11 pesticides), cluster 2 (15 pesticides), cluster 3 (8 pesticides), and cluster 4 (16 pesticides)

The compounds gathered in cluster 4 were mostly herbicides belonging to chloroacetanilides, s-triazines, and substituted ureas (Table 3). They were characterized by high total energy values, whereas polarizability values and the different connectivity indices were low (Fig. 4 and Table A2). Based on these features, cluster 4 was quite close to cluster 1. However, LUMO energy was very high (median LUMO energy value = - 0.119 eV) as compared to cluster 1 (median value = - 0.748 eV). Values of ecotoxicological and environmental parameters placed cluster 4 in between cluster 1 and clusters 2 or 3. It had a low DT50, close to the median value of cluster 1, indicating relatively low persistence of the compounds (median DT50 value = 23 days). The compounds were relatively water-soluble (median Sw = 261 mg/L) and had moderate sorption coefficients (median K_{oc} = 127 L/kg). In terms of ecotoxicity, organisms differed according to the acute or chronic endpoints. Whereas cluster 4 compounds had high acute toxicity for algae and *Lemna* sp. (EC50 = 0.39 µmol/L), they had lower toxicity for invertebrates, *Daphnia* sp., or earthworms than the compounds of cluster 3 (Fig. 3). These results are in accordance with the biochemical effect of herbicides, which disrupt plant physiology and more particularly photosynthesis. Surprisingly, when considering chronic toxicity parameters, the differences in ranking between algae and other organisms (i.e., *Daphnia* sp. and earthworms) were no longer observed. In this cluster, the median BCF value indicated quite a low potential for bioaccumulation along food webs.

Besides well-known relationships between parameters such as water solubility, the octanol-water coefficient, the sorption coefficient, and the bioconcentration factor, the interpretation of the different clusters highlighted certain molecular descriptors likely to have a significant influence on ecotoxicological effects. High dipole moments, high total energy, low connectivity indices, and low MW and surface areas are correlated to low bioaccumulation and low ecotoxicity, as well as to low K_{oc} coefficients and low DT50. On the contrary, low dipole moments, low total energy, high polarizability, high connectivity indices, and high numbers of Cl atoms characterize pesticides with high K_{oc} and medium to high DT50, and high BCF and ecotoxicity for invertebrates. Pesticides combining high total energy, low connectivity indices, and low polarizability have rather high acute ecotoxicity for algae and aquatic plants, but a medium bioaccumulation potential correlated with high water solubility, along with low K_{oc} and DT50 values.

In addition, we can suggest that the relationship between molecular descriptors and environmental behavior and the ensuing consequences on ecotoxicological endpoints partly depend on bioassay type. This should be further investigated because the ecotoxicological effect of pesticides highly depends on the experimental conditions of standardized tests, which generally do not take into account the exposure routes of organisms in natural conditions. To overcome this drawback, one could recommend to conduct ecotoxicological assessment of

pesticides using reproducible laboratory experiments in ecologically relevant conditions. Another issue of concern is the use of standardized tests with a relatively low number of non-target model organisms, which are far from being representative of the possible effects of pesticides on the diversity of non-target organisms.

Conclusion

This proof-of-concept study shows that TyPol allows the classification of pesticides according to their potential ecotoxicological effects on non-target organisms. Despite a relatively low number of pesticides included in the study, the strength of our clustering approach was to consider ecotoxicological endpoints on both terrestrial and aquatic non-target organisms and environmental behavior together. In each cluster, pesticides were gathered according to similar features in terms of environmental behavior and ecotoxicological parameters, related to combinations of values from specific molecular descriptors. This first trial should be extended to a larger number of pesticides and TPs to increase the statistical power in clustering analysis and in the prediction of ecotoxicological parameters.

To do so, one difficulty already encountered in this study is to retrieve relevant ecotoxicological endpoints from public databases: the availability of ecotoxicological parameters in the PPDB or Agritox databases was lower for chronic than for acute toxicity endpoints whatever the targeted organisms. This distortion is due to the assessment of acute toxicity in pesticide regulation by simple and cost- and time-effective tests. It is also well known that measurements of ecotoxicological parameters are far more available for aquatic organisms (invertebrates) than for terrestrial invertebrates. Moreover, interpretation and rationalization of results from terrestrial invertebrates requires a far more complex methodology because tests regularly involve non-equilibrium processes. It also reflects longer time history of regulatory frameworks and European directives for aquatic ecosystems quality than for soils. This emphasizes the need to acquire more ecotoxicological data for a range of pesticides, especially for a range of non-target soil organisms. Filling this gap will (i) improve the possibility to classify pesticides according to their ecotoxicological effects on non-target organisms, and (ii) pave the way for predicting ecotoxicological parameters of pesticides and TPs about which this information is lacking.

Acknowledgements

The authors thank the EcoBASC Network (LabEx BASC, ANR-11-LABX-0034) for financing Harouna Traoré's master's degree grant. We are grateful to Annie Buchwalter for having checked the english on the revised manuscript.

Funding information

This project was supported by the research program “Assessing and reducing environmental risks from plant protection products” funded by the French Ministries in charge of Ecology and Agriculture (IMPEC project).

References

Agritox (2017) http://www.agritox.anses.fr/

Benoit P, Mamy L, Servien R, Li Z, Latrille E, Rossard V, Bessac F, Patureau D., Martin-Laurent F (2017) Categorizing chlordecone potential degradation products to explore their environmental fate. Sci Total Environ 574:781–795

Cronin MTD, Walker JD, Jaworska JS, Comber MHI, Watts CD, Worth AP (2003) Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. Environ Health Perspect 111:1376–1390

Cuhra M, Traavik T, Bøhn T (2013) Clone- and age-dependent toxicity of a glyphosate commercial formulation and its active ingredient in *Daphnia magna.* Ecotoxicology 22(2): 251–262

EFSA Scientific Committee (2016a) Guidance to develop specific protection goals options for environmental risk assessment at EFSA, in relation to biodiversity and ecosystem services. EFSA J 14: 4499. 50 pp. doi:10.2903/j.efsa.2016.4499

EFSA Scientific Committee (2016b) European Food Safety Authority, Scientific opinion on recovery in environmental risk assessments at EFSA. EFSA J 14: 4313. 85 pp. doi:10.2903/j.efsa.2016.4313

EFSA Panel on Plant Protection Products and their Residues (2017) Scientific opinion by Ockleford et al. addressing the state of the science on risk assessment of plant protection products for in-soil organisms. EFSA J 15: 4690. 225 pp. doi:10.2903/j.efsa.2017.4690

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

Hermens J, Balaz S, Damborsky J, Karcher W, Müller M, Peijnenburg W, Sabljic A, Sjöström M (1995) Assessment of QSARs for predicting fate and effects of chemicals in the environment: an international European project. SAR QSAR Environ Res 3:223–236

INERIS, 2017. Portail substances chimiques. https://substances.ineris.fr/fr/search/index.

ISO 11268-1 (1993) Soil quality - Effects of pollutants on earthworms (*Eisenia fetida*) - Part 1: Determination of acute toxicity using artificial soil substrate. International Organization for Standardization, Genève

ISO 11268-2 (1998) Soil Quality - Effects of pollutants on earthworms (*Eisenia fetida*) - Part 2: Determination of effects on reproduction. International Organization for Standardization, Genève

ISO 6341 (1998) Water quality – Determination of acute toxicity of substances to *Daphnia magna*. International Organization for Standardization, Genève

ISO 10706 (2000) Water quality – Determination of long term toxicity of substances to *Daphnia magna*. International Organization for Standardization, Genève

ISO 20079 (2005) Water quality – Determination of toxic effect of water constituents and waste water to duckweed (*Lemna minor*) - duckweed growth inhibition test. International Organization for Standardization, Genève

ISO 8692 (2012) Water quality – Algal growth inhibition test with unicellular green algae. International Organization for Standardization, Genève

Lapenna S, Fuart-Gatnik M, Worth A (2010) Review of QSAR models and software tools for predicting acute and chronic systemic toxicity. JRC Scientific and Technical report, Institute for Health and Consumer Protection EUR 24639 EN, 35 pp.

Mamy L, Patureau D, Barriuso E, Bedos C, Bessac F, Louchart X, Martin-Laurent F, Miege C, Benoit P (2015) Prediction of the fate of organic compounds in the environment from their molecular properties: A review. Crit Rev Environ Sci Tech 45(12):1277–1377

Muir DCG, Howard PH (2006) Are there other persistent organic pollutants? A challenge for environmental chemists. Environ Sci Technol 40:7157–7166

OECD (1984) Earthworms acute toxicity test. Guideline 207. Organisation for Economic Co-operation and Development, Paris, France

OECD (1993) Application of structure-activity relationships to the estimation of properties important in exposure assessment. Environment monographs No 67, Organisation for Economic Co-operation and Development, Paris, France

OECD (1998) *Daphnia magna* reproduction test. Guideline 211. Organisation for Economic Co-operation and Development, Paris, France

OECD (2004) *Daphnia magna* acute immobilisation test. Guideline 202. Organisation for Economic Co-operation and Development, Paris, France

OECD (2006) Freshwater alga and cyanobacteria growth inhibition test. Guidelines 201. Organisation for Economic Co-operation and Development, Paris, France

OECD (2006) *Lemna sp.* growth inhibition test. Guideline 221. Organisation for Economic Co-operation and Development, Paris, France

OECD (2015) Earthworm reproduction test (*Eisenia fetida/Eisenia andrei*). Guideline 222. Paris, France: Organisation for Economic Co-operation and Development, Paris, France

Pavan M, Netzeva TI, Worth AP (2008) Review of literature-based quantitative structure-activity relationship models for bioconcentration. QSAR Comb Sci 27:21-31

PPDB (Pesticide properties database) (2017) http://sitem.herts.ac.uk/aeru/footprint/index2.htm

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)

Russom CL, Breton RL, Walker JD, Bradbury SP (2003) An overview of the use of quantitative structure-activity relationships for ranking and prioritizing large chemical inventories for environmental risk assessments. Environ Tox Chem 22:1810–1821

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

Sabljic A, Nakagawa Y (2014) Biodegradation and quantitative structure-activity relationship (QSAR). In: Chen W, Sabljic A, Cryer SA, Kookana RS (eds) Non-first order degradation and time-dependent sorption of organic chemicals in soil, ACS Symposium Series e-Book, Volume 1174, American Chemical Society, Washington, DC, pp 57-84

Servien R, Mamy L, Li Z, Rossard V, Latrille E, Bessac F, Patureau D, Benoit P (2014) TyPol – A new methodology for organic compounds clustering based on their molecular characteristics and environmental behavior. Chemosphere 111:613–622

Sousa A, Pereira R, Antunes SC, Cachada A, Pereira E, Duarte AC, Gonçalves F (2008) Validation of avoidance assays for the screening assessment of soils under different anthropogenic disturbances. Ecotox Environ Safety 71:661–670

US-EPA EcoTox Knowledgebase (2017) Pesticide Ecotoxicity Database. Environmental Fate and Effects Division of the Office of Pesticide Programs. http://www.ipmcenters.org/Ecotox/DataAccess.cfm

Wold S (1978) Cross-validation estimation of the number of components in factor and principal component analysis. Technometrics 24:397–405

Worrall F (2001) A molecular topology approach to predicting pesticide pollution of groundwater. Environ Sci Technol 35:2282–2287