J Mol Model (2016) 22: 288
DOI 10.1007/s00894-016-3156-0

ORIGINAL PAPER

Acute aquatic toxicity of organic solvents modeled by QSARs

A. Levet[1] · C. Bordes[1] · Y. Clément[1] · P. Mignon[1] · C. Morell[1] · H. Chermette[1] · P. Marote[1] · P. Lantéri[1]

Received: 26 April 2016 / Accepted: 13 October 2016 / Published online: 9 November 2016

Abstract To limit *in vivo* experiments, the use of quantitative structure-activity relationships (QSARs) is advocated by REACH regulation to predict the required fish, invertebrate, and algae EC50 for chemical registration. The aim of this work was to develop reliable QSARs in order to model both invertebrate and algae EC50 for organic solvents, regardless of the mechanism of toxic action involved. EC50 represents the concentration producing the 50 % immobilization of invertebrates or the 50 % growth inhibition of algae. The dataset was composed of 122 organic solvents chemically heterogeneous which were characterized by their invertebrate and/or algae EC50. These solvents were described by physico-chemical descriptors and quantum theoretical parameters calculated via density functional theory. QSAR models were developed by multiple linear regression using the ordinary least squares method and descriptor selection was performed by the Kubinyi function. Invertebrate EC50 was well-described with LogP, dielectric constant, surface tension, and minimal atomic Mulliken charges while algae EC50 of organic solvents (except amines) was predicted with LogP and LUMO energy. To evaluate robustness and predictive performance of the QSARs developed, several strategies have been used to select solvent training sets (random, EC50-based selection and a space-filling design) and both internal and external validations were performed.

Keywords Algae EC50 · DFT · ECOSAR · Ecotoxicity · Invertebrate EC50 · Organic solvents · QSAR

Introduction

Solvents are widely used in many sectors of industry and everyday life such as detergents, agrochemicals, cosmetics, pharmaceuticals, paints, varnishes, inks, etc. Their use has become and more supervised since 2007 by the European regulation Registration, Evaluation, Authorization and restriction of Chemicals (REACH) [1]. Solvents have then to be registered to the European Chemicals Agency (ECHA, http://echa.europa.eu/) requiring, in particular, ecotoxicological information such as acute aquatic toxicity on different trophic levels (fish, invertebrates, and algae). Acute aquatic toxicity is widely evaluated through EC50 values which represent the concentration of chemicals leading to an effect on 50 % of the tested population referred to in a test period [2]. The effect is death for fish, immobility for invertebrates, and inhibition of growth for algae on 96 h, 48 h and 72 h, respectively. To limit *in vivo* testing, alternative methods are advocated by REACH and other organisms like OCDE [3] such as quantitative structure-activity relationships (QSAR). QSARs are mathematical models relating the physico-chemical properties of molecules with their toxicity. QSARs can then be applied after validation to other molecules to predict their activity on the basis that similar chemicals have similar activities [4, 5]. Many QSAR models predict EC50 of fish [6, 7], invertebrates [8, 9], and algae [10, 11]. Various descriptors are involved in such models as LogP [12, 13], hardness [14], HOMO energy [11, 15] or Balaban indice [8]. QSARs are

This paper belongs to Topical Collection Festschrift in Honor of Henry Chermette

Electronic supplementary material The online version of this article (doi:10.1007/s00894-016-3156-0) contains supplementary material, which is available to authorized users.

✉ C. Bordes
bordes@univ-lyon1.fr

[1] Université Claude Bernard Lyon 1, Institut des Sciences Analytiques, Université de Lyon, UMR CNRS 5280, 5 rue de la Doua, 69100 Villeurbanne, France

generally dedicated to specific chemical families as for example benzoic acid [13, 16], alcohols [17], benzene [18], hydrocarbon [19], phenol [20] and may regard or not the toxic action mechanism involved [10]. Only a few general models have been developed to predict invertebrates or algae EC50 with physico-chemical and theoretical parameters for a large set of chemicals [12, 14, 15, 21, 22]. Other models exist based on fragment methods [23–25]. However, to our knowledge, no QSAR was developed for a large dataset of organic solvents.

In a previous study [26], a 4-parameter QSAR was developed allowing the prediction of fish LC50 for organic solvents with the octanol-water partition coefficient, LUMO energy, surface tension, and dielectric constant, regardless of the toxic action mode. The purpose of this study is to complete this previous work by studying EC50 prediction of organic solvents for the two other trophic levels, namely invertebrates and algae. Indeed, the knowledge of EC50 values for fish, invertebrates, and algae allow the determination of the aquatic predicted no-effect concentration (PNEC-aquatic) that represents the concentration below which the exposure to a chemical causes no adverse effects to species in the environment [2]. PNEC-aquatic is of major relevance for environmental risk assessment.

Here, three strategies (namely random selection, EC50-based selection, or space-filling approach) were applied to select the solvents of the training sets used for the QSAR development. The predictive performances of the QSAR models were compared to that of LogP-based relations available in the ECOSAR program (http://www.epa.gov/oppt/newchems/tools/21ecosar.htm).

Material and methods

Data set

The experimental acute toxicity was expressed as pEC50 = log(EC50) for both invertebrates and algae trophic levels which are representative for ecotoxicological evaluation of industrial chemicals. EC50 denotes the concentration in mmol/L producing the 50 % immobilization of invertebrates or the 50 % growth inhibition of algae population referred to in a test period of 48 h and 72 h respectively [2]. For each trophic level, different species were considered: *Daphnia magna* for invertebrates, *Scenedesmus subspicatus* and *Selenastrum capricornutum* for algae. EC50 values were collected in INERIS (http://www.ineris.fr/), ESIS (http://ecb.jrc.ec.europa.eu/esis/) and ECHA (http://echa.europa.eu/) databases for good reliability of the data. When several reliable experimental values were available, the geometric mean was used. The final database includes 154 chemically heterogeneous solvents: 122 solvents were described with invertebrate EC50 values, 75 solvents with algae EC50 values, and 141 with fish LC50 (see Table 1). The trophic level of acute toxicity data found in the literature is specified for each solvent. The pEC50 ranged from −2.67 to 2.73 and from −1.72 to 2.91 for invertebrate and algae trophic level, respectively. Chemicals may be categorized for the three trophic levels as (1) very toxic (EC/LC50 < 1 mg.L^{-1}), (2) toxic (EC/LC50 < 10 mg.L^{-1}), (3) harmful (EC/LC50 < 100 mg.L^{-1}), (4) not harmful (EC/LC50 > 100 mg.L^{-1}) (see Table 1).

Solvent descriptors

Both physico-chemical and quantum theoretical descriptors have been used for solvent description. These 33 descriptors were widely described in a previous paper [26]. Briefly, the physico-chemical descriptors used are those classically employed for both solvent classification [27, 28] and toxicity QSARs [8, 13] purposes: octanol/water partition coefficient (LogP), molecular weight (Mw), boiling point (b_p), density (d), molar volume (V_m), dipole moment (μ), dielectric constant (ε), refractive index (n_r), surface tension (γ), vapor tension (P_{vap}), Hildebrand parameter (δ), and Hansen solubility parameters (δ_d, δ_p, and δ_h). Descriptors values were found in several databases [29–33] (https://reaxys.com; https://scifinder.cas.org).

The ecotoxicity behavior of chemicals may stem from chemical reactivity and selectivity. Therefore, it has been chosen to include relevant quantum descriptors that were computed by the density functional theory (DFT) after geometric optimization with ADF software (http://www.scm.com). Descriptor calculations were performed with the Perdew, Burke, and Ernzerhof (PBE) [34] generalized gradient approximation (GGA) exchange-correlation functional method and a triple zeta (TZP) basis set. These theoretical descriptors were the energy of the highest occupied molecular orbital (HOMO), the energy of the lowest unoccupied molecular orbital (LUMO), the hardness (LUMO-HOMO), the electronegativity ((HOMO + LUMO)/2), the polarisability (α), the maximal (q_{max}) and minimal (q_{min}) atomic Mulliken charges, the maximal (V_{max}) and minimal (V_{min}) electrostatic potential values, the surface area (Surf), the molecular volume (Vol), and the molecule's ovality (Ov). Three descriptors were added and are related to the electrostatic potential (ESP) computed between -3 eV and +3 eV on the solvent accessible surface around the molecule: the surface Sneg with ESP lower than −0.1 eV, the surface S with ESP ranging from −0.1 eV up to 0.1 eV, and the surface Spos with ESP larger than 0.1 eV. S represents hydrophobic regions while Sneg and Spos represent hydrophilic ones in relation with base or acid Lewis nature. Topological descriptors were also calculated using ProChemist software (http://pro.chemist.online.fr/) such

Table 1 List of the organic solvents in the dataset with the corresponding type of acute toxicity data (F for fish, I for invertebrates and A for algae) and the ecotoxicological class for the three trophic levels assigned according to the European Commission (EC, 1991): (1) very toxic (LC50 < 1 mg.L-1), (2) toxic (LC50 < 10 mg.L-1), (3) harmful (LC50 < 100 mg.L-1), (4) not harmful (LC50 > 100 mg.L-1)

CAS number	Name	Molecular formula	EC50	Ecotox class fish	Ecotox class inv.	Ecotox class algae
56-23-5	Carbon tetrachloride	CCl_4	F,I,A	3	3	3
56-81-5	Glycerol	$C_3H_8O_3$	F,I	4	4	
57-55-6	1,2-Propylene glycol	$C_3H_8O_2$	F,I,A	4	4	4
60-29-7	Diethyl ether	$C_4H_{10}O$	F,I	4	4	
62-53-3	Aniline	C_6H_7N	F,I,A	3	1	4
64-17-5	Ethanol	C_2H_6O	F,I	4	4	
64-18-6	Formic acid	CH_2O_2	F,I	4	4	
64-19-7	Acetic acid	$C_2H_4O_2$	F	4		
67-56-1	Methanol	CH_4O	F,I,A	4	4	4
67-63-0	Isopropyl alcohol	C_3H_8O	F,I	4	4	
67-64-1	Acetone	C_3H_6O	F,I	4	4	
67-66-3	Chloroform	$CHCl_3$	F,I,A	3	3	3
67-68-5	Dimethyl sulfoxide	C_2H_6OS	F,I,A	4	4	4
68-12-2	N,N-Dimethylformamide	C_3H_7NO	F,I	4	4	
71-23-8	1-Propanol	C_3H_8O	F,I	4	4	
71-36-3	1-Butanol	$C_4H_{10}O$	F,I	4	4	
71-41-0	1-Pentanol	$C_5H_{12}O$	F,I,A	4	4	4
71-43-2	Benzene	C_6H_6	F,I,A	3	2	3
71-55-6	1,1,1-Trichloroethane	$C_2H_3Cl_3$	F,I,A	3	3	3
74-95-3	Dibromomethane	CH_2Br_2	F,I,A	3	3	4
75-05-8	Acetonitrile	C_2H_3N	F,I,A	4	4	4
75-09-2	Dichloromethane	CH_2Cl_2	F,I	4	4	
75-12-7	Formamide	CH_3NO	F	4		
75-15-0	Carbon disulfide	CS_2	F,I	2	2	
75-29-6	2-Chloropropane	C_3H_7Cl	F	4		
75-34-3	1,1-Dichloroethane	$C_2H_4Cl_2$	F,I	4	3	
75-35-4	1,1-Dichloroethylene	$C_2H_2Cl_2$	F,I,A	4	3	2
75-65-0	tert-Butanol	$C_4H_{10}O$	F,A	4		4
75-89-8	2,2,2-Trifluoroethanol	$C_2H_3F_3O$	F	4		
75-97-8	3,3-Dimethyl-2-butanone	$C_6H_{12}O$	F	4		
76-05-1	Trifluoroacetic acid	$C_2HO_2F_3$	I,A		4	4
78-59-1	Isophorone	$C_9H_{14}O$	F,I,A	4	4	4
78-83-1	Isobutanol	$C_4H_{10}O$	F,I,A	4	4	4
78-92-2	2-Butanol	$C_4H_{10}O$	F,I	4	4	
78-93-3	Methyl ethyl ketone	C_4H_8O	F,I	4	4	
79-00-5	1,1,2-Trichloroethane	$C_2H_3Cl_3$	F,I,A	3	3	4
79-01-6	1,1,2-Trichloroethylene	C_2HCl_3	F,I,A	3	3	3
79-16-3	N-Methylacetamide	C_3H_7NO	F,I	4	4	
79-20-9	Methyl acetate	$C_3H_6O_2$	F,I	4	4	
79-34-5	1,1,2,2-Tetrachloroethane	$C_2H_2Cl_4$	F,I	3	3	
91-17-8	Decahydronaphtalene	$C_{10}H_{18}$	F,I	2	2	
91-22-5	Quinoline	C_9H_7N	F,I,A	3	3	2
95-47-6	o-Xylene	C_8H_{10}	F,I	2	2	
95-48-7	o-Cresol	C_7H_8O	F,I,A	2	3	4
95-50-1	o-Dichlorobenzene	$C_6H_4Cl_2$	F,I	2	2	
96-22-0	3-Pentanone	$C_5H_{10}O$	F	4		
96-49-1	Ethylene carbonate	$C_3H_4O_3$	F,I	4	4	
97-99-4	Tetrahydrofurfuryl alcohol	$C_5H_{10}O_2$	F	4		
98-00-0	Furfuryl alcohol	$C_5H_6O_2$	F	4		

Table 1 (continued)

CAS number	Name	Molecular formula	EC50	Ecotox class fish	Ecotox class inv.	Ecotox class algae
98-01-1	Furfural	$C_5H_4O_2$	F,I	3	3	
98-82-8	Isopropylbenzene	C_9H_{12}	F,I,A	2	2	2
98-86-2	Acetophenone	C_8H_8O	F,I,A	4	4	3
98-95-3	Nitrobenzene	$C_6H_5NO_2$	F,I,A	3	3	3
100-41-4	Ethylbenzene	C_8H_{10}	I,A		2	2
100-42-5	Styrene	C_8H_8	F,I,A	2	2	2
100-47-0	Benzonitrile	C_7H_5N	F	4		
100-51-6	Benzyl alcohol	C_7H_8O	F,I,A	4	4	4
100-52-7	Benzaldehyde	C_7H_6O	F,I,A	2	3	4
101-84-8	Diphenyl ether	$C_{12}H_{10}O$	F,I,A	2	2	2
102-82-9	Tributylamine	$C_{12}H_{27}N$	F,I,A	3	2	2
103-50-4	Dibenzyl ether	$C_{14}H_{14}O$	F,I,A	2	1	2
103-73-1	Phenetole	$C_8H_{10}O$	F	3		
104-51-8	Butylbenzene	$C_{10}H_{14}$	F	2		
104-76-7	2-Ethylhexanol	$C_8H_{18}O$	F,I,A	3	3	3
105-37-3	Ethyl propionate	$C_5H_{10}O_2$	F	4		
106-42-3	p-Xylene	C_8H_{10}	F,I,A	2	2	2
107-06-2	1,2-Dichloroethane	$C_2H_4Cl_2$	F,I,A	4	4	4
107-07-3	2-Chloroethanol	C_2H_5ClO	F,I	3	4	
107-10-8	Propylamine	C_3H_9N	I		3	
107-15-3	1,2-Diaminoethane	$C_2H_8N_2$	F,I,A	4	3	4
107-21-1	Ethylene glycol	$C_2H_6O_2$	F,I,A	4	4	4
107-31-3	Methyl formate	$C_2H_4O_2$	F,I,A	4	4	4
107-88-0	1,3-Butanediol	$C_4H_{10}O_2$	F,I,A	4	4	4
108-03-2	1-Nitropropane	$C_3H_7NO_2$	F,I,A	4	4	3
108-10-1	Methyl isobutyl ketone	$C_6H_{12}O$	F,I	4	4	
108-20-3	Diisopropyl ether	$C_6H_{14}O$	F,I	4	4	
108-21-4	Isopropyl acetate	$C_5H_{10}O_2$	F,I	4	4	
108-24-7	Acetic acid, anhydride	$C_4H_6O_3$	F,I	4	3	
108-38-3	m-Xylene	C_8H_{10}	F,I,A	3	2	2
108-39-4	m-Cresol	C_7H_8O	F,I,A	3	3	4
108-67-8	Mesitylene	C_9H_{12}	F,I	2	2	
108-83-8	Diisobutyl ketone	$C_9H_{18}O$	F,I,A	3	4	3
108-87-2	Methylcyclohexane	C_7H_{14}	F	3		
108-88-3	Toluene	C_7H_8	F,I	2	3	
108-89-4	4-Methylpyridine	C_6H_7N	F	4		
108-90-7	Monochlorobenzene	C_6H_5Cl	F,I,A	2	2	3
108-93-0	Cyclohexanol	$C_6H_{12}O$	F,I,A	4	3	3
108-94-1	Cyclohexanone	$C_6H_{10}O$	F,I,A	4	4	3
108-95-2	Phenol	C_6H_6O	F,I	2	3	
108-99-6	3-Methylpyridine	C_6H_7N	F,I,A	4	4	4
109-60-4	n-Propyl acetate	$C_5H_{10}O_2$	F,I,A	3	3	4
109-65-9	1-Bromobutane	C_4H_9Br	F	3		
109-66-0	n-Pentane	C_5H_{12}	F,I,A	2	2	2
109-69-3	1-Chlorobutane	C_4H_9Cl	F,A	4		4
109-73-9	n-Butylamine	$C_4H_{11}N$	F,I	3	2	
109-86-4	2-Methoxyethanol	$C_3H_8O_2$	A			4
109-87-5	Dimethoxymethane	$C_3H_8O_2$	F,I,A	4	3	3
109-89-7	Diethylamine	$C_4H_{11}N$	F	4		

Table 1 (continued)

CAS number	Name	Molecular formula	EC50	Ecotox class fish	Ecotox class inv.	Ecotox class algae
109-94-4	Ethyl formate	$C_3H_6O_2$	F,I	4	4	
109-99-9	Tetrahydrofuran	C_4H_8O	F	4		
110-12-3	5-Methyl-2-hexanone	$C_7H_{14}O$	F,I,A	4	3	4
110-19-0	Isobutyl acetate	$C_6H_{12}O_2$	F,I	3	4	
110-63-4	1,4-Butanediol	$C_4H_{10}O_2$	I,A		4	4
110-71-4	1,2-Dimethoxyethane	$C_4H_{10}O_2$	I,A		4	4
110-80-5	2-Ethoxyethanol	$C_4H_{10}O_2$	I		4	
110-82-7	Cyclohexane	C_6H_{12}	F,I,A	2	2	2
110-83-8	Cyclohexene	C_6H_{10}	F,I,A	2	2	3
110-86-1	Pyridine	C_5H_5N	F,I,A	4	4	4
110-91-8	Morpholine	C_4H_9NO	F,I,A	4	3	3
111-27-3	1-Hexanol	$C_6H_{14}O$	F,I,A	4	4	3
111-40-0	Diethylenetriamine	$C_4H_{13}N_3$	F,I	4	3	
111-46-6	Diethylene glycol	$C_4H_{10}O_3$	F,I,A	4	4	4
111-76-2	2-Butoxyethanol	$C_6H_{14}O_2$	F,I,A	4	4	4
111-77-3	Diethylene glycol monomethyl ether	$C_5H_{12}O_3$	F,I	4	4	
111-87-5	1-Octanol	$C_8H_{18}O$	F,I	3	2	
111-90-0	Diethylene glycol monoethyl ether	$C_6H_{14}O_3$	I		4	
111-96-6	Diethylene glycol dimethyl ether	$C_6H_{14}O_3$	F,I	4	4	
112-25-4	Ethylene glycol hexyl ether	$C_8H_{18}O_2$	F,I,A	4	4	4
112-27-6	Triethylene glycol	$C_6H_{14}O_4$	F,I	4	4	
112-60-7	Tetraethylene glycol	$C_8H_{18}O_5$	I		4	
119-64-2	Tetralin	$C_{10}H_{12}$	F,I,A	2	2	2
120-82-1	1,2,4-Trichlorobenzene	$C_6H_3Cl_3$	F,I,A	2	2	2
121-44-8	Triethylamine	$C_6H_{15}N$	F,I	3	3	
121-69-7	N,N-Dimethylaniline	$C_8H_{11}N$	F,I	3	2	
123-51-3	3-Methyl-1-butanol	$C_5H_{12}O$	F	4		
123-54-6	Acetyl acetone	$C_5H_8O_2$	F,I	3	3	
123-75-1	Pyrrolidine	C_4H_9N	F	4		
123-86-4	n-Butyl acetate	$C_6H_{12}O_2$	F,I,A	3	3	4
123-91-1	1,4-Dioxane	$C_4H_8O_2$	F,I	4	4	
126-33-0	Sulfolane	$C_4H_8O_2S$	F,I,A	4	4	4
127-18-4	1,1,2,2-Tetrachloroethylene	C_2Cl_4	F,I,A	3	3	2
131-11-3	Dimethyl phthalate	$C_{10}H_{10}O_4$	F,A	3		4
137-32-6	2-Methyl-1-butanol	$C_5H_{12}O$	F,I,A	3	4	4
141-43-5	Ethanolamine	C_2H_7NO	I		3	
141-78-6	Ethyl acetate	$C_4H_8O_2$	F,I	4	4	
141-79-7	Methyl isobutyl ketone	$C_6H_{10}O$	F,I	3	3	
142-96-1	Dibutyl ether	$C_8H_{18}O$	F,I	3	3	
156-59-2	1,2-Dichloroéthylène	$C_2H_2Cl_2$	F	3		
287-92-3	Cyclopentane	C_5H_{10}	F,I	2	3	
462-06-6	Fluorobenzene	C_6H_5F	F	2		
504-63-2	1,3-Propylene glycol	$C_3H_8O_2$	F,A	4		4
540-54-5	Propyl chloride	C_3H_7Cl	F	3		
540-84-1	2,2,4-Trimethyl pentane	C_8H_{18}	I		1	
541-73-1	m-Dichlorobenzene	$C_6H_4Cl_2$	F,I,A	2	2	2
563-80-4	Methyl isopropyl ketone	$C_5H_{10}O$	A			3

Table 1 (continued)

CAS number	Name	Molecular formula	EC50	Ecotox class fish	Ecotox class inv.	Ecotox class algae
565-80-0	2,4-Dimethyl-3-pentanone	$C_7H_{14}O$	F	4		
616-45-5	2-Pyrrolidone	C_4H_7NO	F,A	4		4
628-63-7	n-Pentyl acetate	$C_7H_{14}O_2$	F,I	3	3	
872-50-4	N-Methylpyrrolidone	C_5H_9NO	I		4	
1119-40-0	Dimethyl glutarate	$C_7H_{12}O_4$	F,I	3	4	
1634-04-4	Methyl tert-butyl ether	$C_5H_{12}O$	F,I,A	4	4	4
2807-30-9	Ethylene glycol monopropyl ether	$C_5H_{12}O_2$	F,A	4		4
5989-27-5	d-Limonene	$C_{10}H_{16}$	F,I,A	1	1	4
29911-28-2	Di(propylene glycol)butyl ether	$C_{10}H_{22}O_3$	F	4		

as Wiener [35], Balaban [36], Randic [37], Kier Indices [38].

Model development procedure

Descriptor selection

QSAR models were developed by multiple linear regression (MLR) using the ordinary least squares method. The use of this simple learning method makes such models easy to interpret and to apply. MLR calculations were achieved by using the enhanced replacement method (ERM) with Matlab 7.9 software (QSAR/QSPR search algorithms Toolbox; www.mathworks.fr/products/matlab/). ERM algorithm requires a smaller number of linear regressions than a time-consuming Full Search method while obtaining identical results [39]. Descriptor selection was performed by the Kubinyi function (FIT) [40], expressed as:

$$FIT = \frac{R^2(N-d-1)}{\left(N+d^2\right)\left(1-R^2\right)} \tag{1}$$

where R^2 is the determination coefficient, d the number of descriptors selected in the model and N the number of solvents in the training set. The optimal number of descriptors selected in the model d_{opt} corresponds to the maximum value of FIT in the plot FIT vs d. The FIT statistical parameter is preferred to the Fisher ratio F too sensitive to changes in small d values and poorly sensitive to changes in large d values. The choice of the descriptors was confirmed by performing Student's *t*-test at a confidence level of 95 %.

Model validation

All the models developed were evaluated through the determination coefficient R^2, the adjusted determination coefficient R^2_a, and the mean absolute error or mean residual (MAE). The predictive power and robustness of the models developed were assessed by internal and external validation techniques.

Fivefold cross-validation was employed for model internal validation. The training set was randomly divided into five subsets of approximately equal size. Four subsets were used as the training set and the last one as the test set. This procedure was repeated five times so that every subset is selected as a test set once. The squared cross-validated correlation coefficient R^2_{CV} was computed.

In order to check that developed QSARs did not depend on a particular distribution of solvents and according to a previous work [26], three strategies have been used to split the solvent dataset into training and test sets for external validation purpose. Briefly a random selection (1) of training and test sets was performed and repeated five times with a four to one size ratio. A Y-based selection (2) was used to obtain a good representation of all types of solvents in terms of ecotoxicity profile. This strategy classified solvents by ascending pEC50 and to kept three out of four solvents in the training set. To select a training set representative of the solvent space studied, a space filling (SF) technique (3) based on minimax distance criterion was used [41] (cover.design function, fields package, R software (www.r-project.org/main.shtml/)). The profile of the candidate solvents was defined through DFT descriptors as well as LogP well-known to be the relevant characteristic for ecotoxicity explanation. To reduce solvent profile dimensions, principal component analysis (PCA) was first performed and solvents were described by their scores on the principal components explaining 90 % of the dataset variance before to be screened with the SF algorithm. The solution set was determined after 20 runs for convergence and quality purpose.

As reviewed by Chirico and Gramatica [42, 43], several external validation criteria may be used to assess QSAR predictivity and robustness: predictive squared correlation coefficients such as Q^2_{F1} [44], Q^2_{F2} [45], Q^2_{F3} [46, 47] and other criteria such as r^2_m [48, 49], the Golbraikh-Tropsha method

[50] or the concordance correlation coefficient [42, 51]. Here, the classical squared correlation coefficient Q^2_{F1} which is advocated in the OECD guidelines [52] was used and is expressed as:

$$Q^2_{F1} = 1 - \frac{\sum_{i=1}^{n_{test}} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n_{test}} \left(y_i - \overline{y}_{train}\right)^2} \tag{2}$$

where y_i and y_i are the observed and the predicted log(EC50) respectively of the test set solvents. $\overline{y}_{train}$ is the mean observed log(EC50) of the training set solvents. To well-assess the external predictivity of the QSARs developed, a second external validation criteria was also evaluated, the concordance correlation coefficient (CCC) expressed as:

$$CCC = \frac{2\sum_{i=1}^{n_{test}} \left(y_i - \overline{y}\right)\left(\hat{y} - \overline{\hat{y}}\right)}{\sum_{i=1}^{n_{test}} \left(y_i - \overline{y}\right)^2 + \sum_{i=1}^{n_{test}} \left(\hat{y}_i - \overline{\hat{y}}\right)^2 + n_{test}\left(\overline{y} - \overline{\hat{y}}\right)^2} \tag{3}$$

where $\overline{y}$ and $\overline{\hat{y}}$ are the mean observed and the mean predicted log(EC50) respectively of the test set solvents. The CCC criteria was chosen since Chirico and Gramatica [43] demonstrated its high reliability (compared to the other validation criteria) by studying the predictivity of QSARs developed from simulated data with different levels of bias. As recommended by these authors, the used acceptance values of Q^2_{F1} and CCC were 0.70 and 0.85, respectively.

Applicability domain

Developed QSARs also require the definition of the corresponding applicability domain (AD) for estimating the reliability in the prediction of a new molecule [53]. Predicted activity for only those compounds that fall into this domain may be considered reliable. AD may be determined by using several approaches [54]. Here, we used a common one based on the leverage values for each chemical [55] which are calculated as follows:

$$h_i = x_i^T \left(X^T X\right)^{-1} x_i \tag{4}$$

where x_i is the descriptor vector of the chemical i and X the model matrix derived from the training set descriptor values. A warning leverage h* is defined and expressed as $h^* = \frac{3p}{n_{train}}$ where p is the number of model parameters [54]. A chemical belonging to the training set with both $h_i > h^*$ and small standardized residuals (smaller than a value of 3 corresponding to 99 % of the normally distributed data) would reinforce the model; while high leverage compounds with large standardized residuals are expected to badly influence the model. The representation of the cross-validated standardized residuals vs the compound leverages is called the Williams plot and allows detecting both the response outliers and the structurally influential chemicals of a model.

ECOSAR

Several QSAR-based programs are available to predict the ecotoxicological risks associated with chemicals, such as ECOSAR, TOPKAT, DEREK, MCASE, ADAPT, etc. [56, 57]. ECOSAR is a freely available program which was developed on experimental data by the United States Environmental Protection Agency (US EPA). Moreover, due to its good predictive performances [58, 59], ECOSAR is widely used for chemicals risk assessment [10, 16, 58–62]. Thus, in this work, the QSARs implemented in ECOSAR which are only based on LogP value and dedicated to neutral organics were used to predict invertebrate or algae pEC50. Neutral organic compounds correspond to alcohols, acetals, ketones, ethers, alkyl halides, aryl halides, aromatic hydrocarbons, halogenated aromatic hydrocarbons, halogenated aliphatic hydrocarbons, sulfides, and disulfides.

For invertebrate (Daphnia), the relation between EC50 and LogP is developed for 115 neutral organics compounds ($R^2 = 0.771$):

$$\text{Log 48-h EC50 Invertebrate (mmol/L)} \\ = -0.8157 \text{ logP} + 1.2695 \tag{5}$$

For algae, the relation between EC50 and LogP is developed for 51 neutral organics compounds ($R^2 = 0.596$):

$$\text{Log 96-h EC50 Algae (mmol/L)} \\ = -0.6271 \log P + 0.5687 \tag{6}$$

Results and discussion

Data set characterization

The three Hansen solubility parameters were used to summarize the chemical profile of the solvents for graphic representation purpose. In order to observe the solvent space studied in a relatively reliable 2D-representation, a PCA was performed by using the Hansen descriptors as initial variables (Figure 1a). As observed in Fig. 1b, PC1 was related both to polarity and hydrogen bonding parameters while PC2 reflected the disperse part of the solubility parameter. The solvent dataset was chemically heterogeneous as also shown

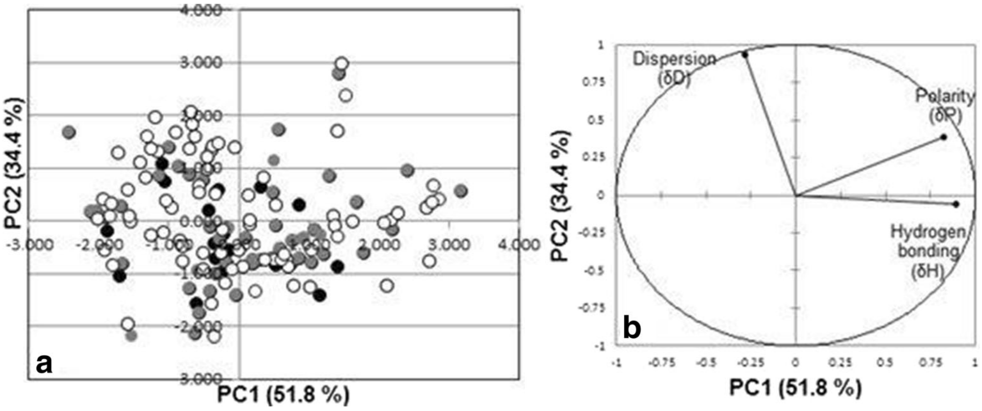

Fig. 1 Representation of the solvents characterized in the data set by fish LC50 (●), invertebrate EC50 (●), and/or algae EC50 (○) in the PC1/PC2 score plot determined by PCA from the three Hansen solubility parameters (**a**) and the corresponding loading plot (**b**). PC1 explains 51.8 % of the variation and the second component PC2 34.4 %

by the large coverage of Hansen solubility space in Fig. 1a. Moreover, this representation confirms that the whole dataset is well-characterized by LC or EC50 for each trophic level: fish [26], invertebrates, and algae. The chi-square test (with the significance level 0.05) indicated that the distributions of both invertebrates and algae EC50 (Fig. 2) conform closely to the normal distribution: the corresponding p-values were 0.401 and 0.175, respectively. Invertebrate logEC50 (48 h) ranged from −2.7 to 2.7 while Algae Log EC50 ranged from −1.7 to 2.9.

Comparisons of toxicity of trophic levels

Figure 3 shows the toxicity correlation between each pair of trophic levels. As observed by many authors [21, 63, 64], the highest correlation ($R^2 = 0.777$) was obtained between fish and *Daphnia magna* transducing their similar sensitivity to organic solvents (Fig. 3a). Aniline was highlighted since its toxicity is much higher toward invertebrates than fish (pEC50 = −2.67 and −0.33 respectively). Aniline and derivatives are considered to be narcotics for fish and more toxic to invertebrates [65]. Similar correlations were obtained between the toxicity values of algae and fish ($R^2 = 0.580$) and algae and

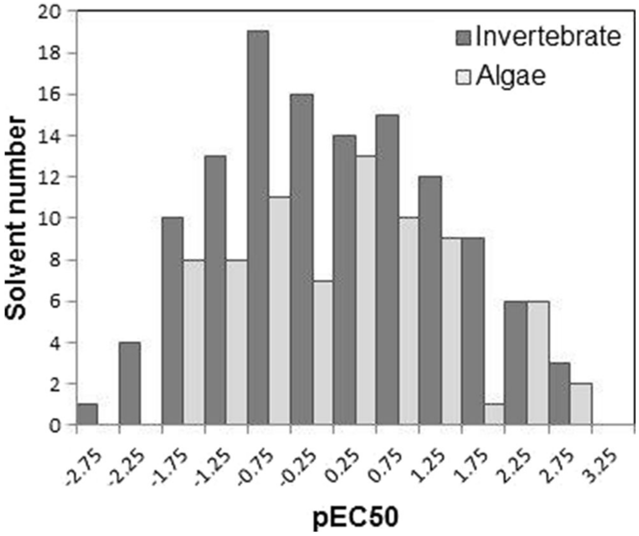

Fig. 2 Histograms of both invertebrates and algae pEC50

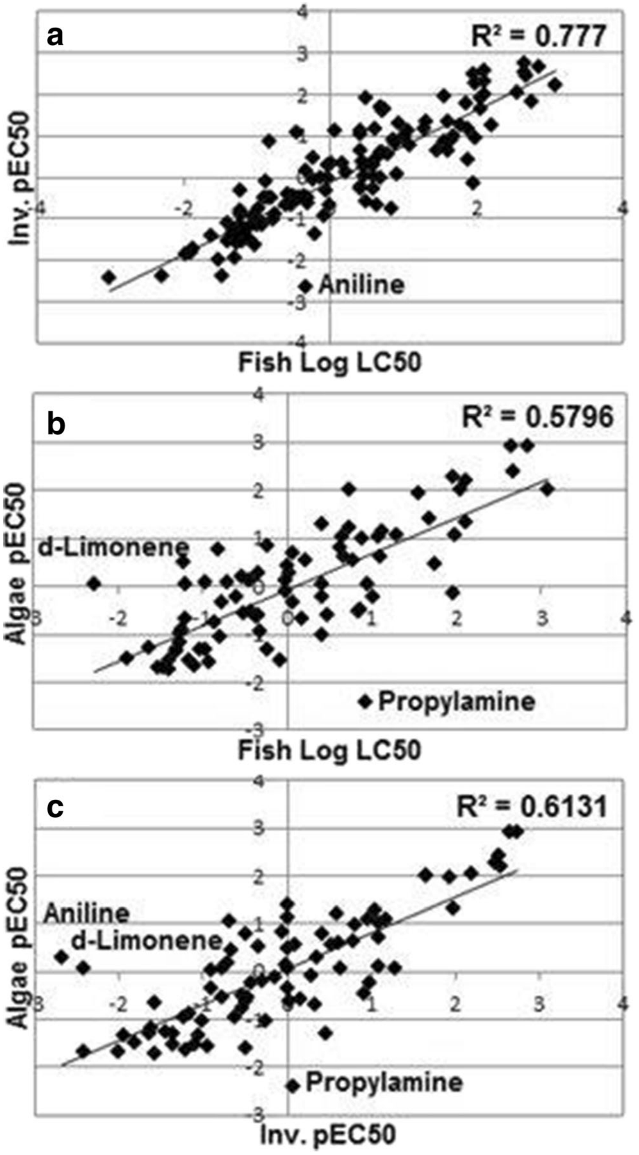

Fig. 3 Toxicity of invertebrates vs fish (**a**), algae vs ish (**b**), and algae vs invertebrates (**c**)

Daphnia magna ($R^2 = 0.613$) (Fig. 3b and c). These results are in agreement with several studies dedicated to interspecies toxicity correlations [21, 64, 66, 67]. Amine compounds (propylamine, tributylamine, n-butylamine, and ethanolamine) toxicity was much higher toward algae than fish or invertebrates as observed by Christensen et al. [68] or Escher et al. [69] while the opposite effect was observed for the most lipophilic solvent of the dataset d-limonene. For the solvent classes large enough (halogenated compounds, ether and orthoesters, acyl compounds and alcohols), we were not able to highlight species more sensitive or less sensitive than the others. The toxicity of the ten aromatic hydrocarbons studied was high toward the three trophic levels.

QSAR development for invertebrate pEC50

Descriptor selection

The descriptors were selected from each training set by multiple linear regression (ERM) with FIT criterion which allows a good compromise between the number of descriptors selected in the model and model fitting quality. The maximum values of FIT corresponded to a 4 or 5-parameter QSAR for invertebrate pEC50 prediction, depending on the strategy of training set selection. As expected, LogP was included in all the models developed and was associated with the surface tension and the minimal atomic Mulliken charges (q_{min}). HOMO energy and/or dielectric constant were also selected. The space-filling design approach allowed the selection of training solvents well-representative of the whole dataset with a good coverage of the solvent space. This strategy led to a 4-descriptors model involving LogP, ε, γ, and q_{min}. From the EC50-based training set, these 4 descriptors were associated with the HOMO energy.

Validation

QSAR including LogP, ε, γ, and q_{min} as explanatory variables (see Table 1 in Supplementary material) led to the best regression performances. This model was externally validated. Results were quite similar for all training sets regardless of the selection strategy: the determination coefficients (R^2_{train}, $R^2_{A,train}$) ranged from 0.689 up to 0.752 and MAE_{train} between 0.512 and 0.535 (see Table 2). For test sets, regression quality was satisfactory ($Q^2_{F1} > 0.7$ and $CCC > 0.85$), especially with SF design ($Q^2_{F1} = 0.864$, $MAE_{test} = 0.425$ and $CCC = 0.907$). The QSAR developed was also internally validated with five-fold cross-validation and a corresponding determination coefficient R^2_{CV} =0.704 in good agreement with the model.

The best QSAR for invertebrate pEC50 prediction was the following:

$$Log\ 48\text{–}h\ Inv.\ EC50\ (mmol/L) = 1.276(\pm 0.418) - 0.480(\pm 0.061)\ LogP - 0.048(\pm 0.011)\gamma + 0.027(\pm 0.007)\ \varepsilon - 0.951(\pm 0.454)\ q_{min} \tag{7}$$

Figure 4a shows the predicted vs experimental invertebrate pEC50 both for training and test sets. In a previous work [26], we already highlighted the significance of LogP, ε, and γ to predict the fish LC50 of organic solvents. Compared to the QSAR developed for invertebrate pEC50, only the fourth parameter differed with the LUMO energy selected instead of q_{min}. These similar results corroborate that fish and daphnids have similar sensitivity to chemicals and especially to narcotic compounds which represented the major part of the solvent dataset. Polar or non-polar narcosis involves non-specific non-covalent interactions with membranes leading to their disruption. This baseline toxicity mechanism is essentially governed

Table 2 Training and test sets characteristics for the three selection strategies used (defined by using (1) random, (2) EC50-based, and (3) space-filling selections)

Training selection	Training solvents		Test solvents		Inv. QSAR including LogP; ε; γ; q_{min}						Algae QSAR including LogP and LUMO					
	Inv.	**Alg.**	**Inv.**	**Alg.**	$\mathbf{R^2_{train}}$	$\mathbf{R^2_{A,train}}$	$\mathbf{MAE_{train}}$	$\mathbf{MAE_{test}}$	$\mathbf{Q^2_{F1}}$	**CCC**	$\mathbf{R^2_{train}}$	$\mathbf{R^2_{A,train}}$	$\mathbf{MAE_{train}}$	$\mathbf{MAE_{test}}$	$\mathbf{Q^2_{F1}}$	**CCC**
Random (1)	92	51	30	20	0.752	0.740	0.522	0.517	0.716	0.852	0.731	0.720	0.473	0.427	0.770	0.874
EC50-based (2)	92	55	30	16	0.744	0.732	0.512	0.557	0.712	0.863	0.744	0.734	0.457	0.447	0.729	0.887
Space filling (3)	90	51	32	20	0.703	0.689	0.535	0.425	0.864	0.907	0.718	0.706	0.466	0.513	0.809	0.869

The best QSARs for both invertebrates and algae pEC50 are described by the determination coefficients and the mean absolute errors corresponding to both the training and test sets. The validation criteria Q^2_{F1} and CCC are also reported

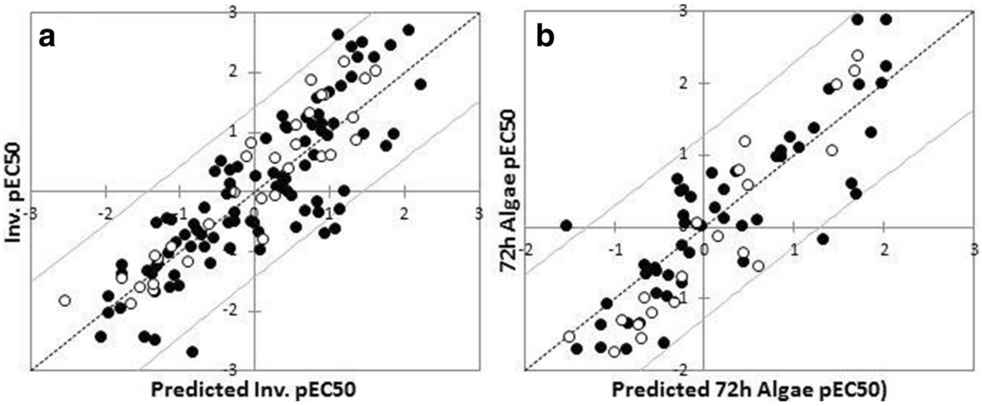

Fig. 4 Predicted vs experimental p(EC50) of invertebrates (**a**) and algae (**b**) for training (●) and test (○) solvents

by LogP so many QSARs (as in ECOSAR program) solely used LogP as a single descriptor. As expected and as observed by many authors [14, 24, 70], the negative coefficient of LogP shows that lipophilic chemicals will have more toxic effects than hydrophilic ones since they are more bioavalaible (transmembrane passage is promoted) and bioaccumulative. The dielectric constant (ε) shows the ability of a solvent for charge separation and provides a rough measure of its polarity. This parameters appeared to be relevant for daphnids toxicity explanation with a positive coefficient. The negative contribution of γ indicates that high surface tension would increase chemicals toxicity which may be explained by a promoted membrane penetration. The last selected descriptor was the minimal atomic Mulliken charges (q_{min}) which translate the ability of the solvent to function as an electron-pair donor (Lewis base). Low q_{min} would promote toxicity. One explanation could be that H_2O would strongly solvate Lewis bases limiting their reactivity and then their ecotoxicity. Faucon et al. [14] showed that ecotoxicity of 96 heterogeneous chemicals toward daphnids increase with the decrease of both LogP and the electronic descriptor hardness (absolute value). They explained the latter behavior by a less favorable solvation of soft compounds which became more reactive and toxic than hard basis or oxygen anions.

To visualize the applicability domain of the developed QSAR, the William's plot was represented in Fig. 5a. From this plot, the AD is established inside a squared area within ±3 standard deviations and a leverage threshold h* of 0.13. No chemical of the test set exceed the warning leverage h* indicating that their predicted activity can be considered as reliable as those of the training chemicals. The activity of four molecules belonging to the training set was less well predicted but remained satisfactory (standardized residuals < 3). According to fish pLC50 prediction [26], the strongest positive residual was obtained for a glycol solvent (diethylene glycol) which exhibits very high 48 h Inv. EC50 (48,900 mg/L). As Papa et al. [71], we expected that the high EC50 values should be difficult to precisely measure.

The solvents which were the most underestimated by the QSAR developed were amine compounds: aniline, diethylenetriamine, and 1,2-diaminoethane. Similar results have been observed for fish pLC50 prediction of organic solvents [26]. Amine solvents may exhibit toxic action in excess of narcosis baseline through specific mechanisms called amine narcosis [65]. Figure 5a also shows that ethylene carbonate has a leverage value 0.41 greater than h* with a small standardized residual. This solvent may stabilize the model and make it more accurate.

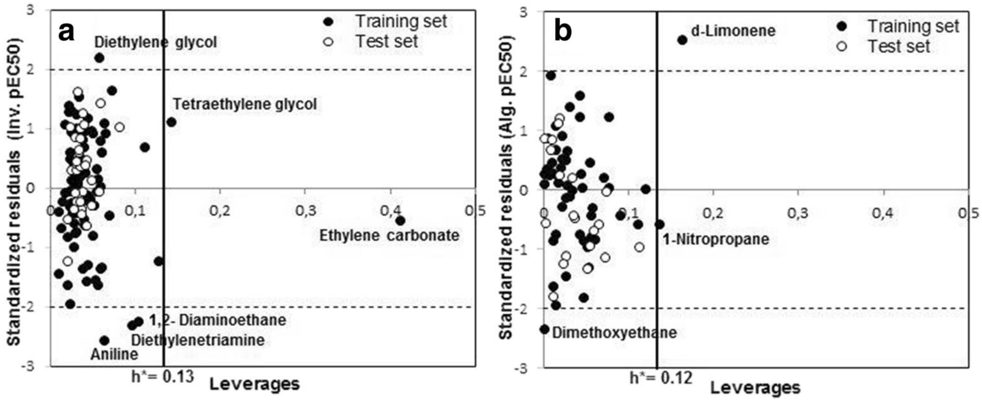

Fig. 5 Williams plot for the QSAR model of invertebrate (**a**) and algae (**b**) pEC50

Finally, the QSAR (Eq. 6) implemented in ECOSAR program was used to predict Inv. pEC50 of the 85 neutral organic solvents contained in the dataset. A comparison with the performance of the developed 4-parameter model (Eq. 7) showed similar results for these compounds with $R^2 = 0.745$ and MAE = 0.497 for ECOSAR relation and $R^2 = 0.758$ and MAE = 0.482 for Eq. 5. However, Eq. 7 was relatively robust since its predictive power for Inv. pEC50 of the remaining 37 reactive or ionizable solvents of the dataset led to $R^2 = 0.639$ (MAE = 0.675) against $R^2 = 0.135$ (MAE = 0.860) for ECOSAR model.

QSAR development for Algae pEC50

As for fish and invertebrate pEC50 prediction, we tried to develop a QSAR for algae pEC50 modeling regardless of the mechanism of toxic action involved. No satisfying models (R^2 around 0.5) were obtained essentially due to the presence of amine solvents in the dataset. Therefore, we chose to remove from the initial dataset all the amine solvents namely ethanolamine, propylamine, n-butylamine, tributylamine, 1,2-diaminoethane, morpholine, and aniline. Amine solvents are well-known to be highly toxic toward algae and this behavior is often related to a pH-dependent toxicity [68, 69]. Neuwoehner and Escher [72] suggest the high toxicity of aliphatic amines in algae is due to a toxicokinetic effect induced by their speciation and not to a specific mechanism of toxic action. Aliphatic amines speciation would be different in the external medium compared to the algae cell in which pH remains independent of external pH.

Space-filling and EC50-based selections led on the basis of FIT criterion to a 2-parameter QSAR for algae EC50 prediction involving LogP and LUMO energy (see Table 1 in Supplementary material). QSAR models with only these two descriptors were already found by several authors [20, 73, 74].

The 2-parameter QSAR was externally validated showing robust results for each selection strategy: the determination coefficients (R^2_{train}, $R^2_{A,train}$) ranged from 0.706 up to 0.744 and MAE from 0.427 up to 0.513 for both training and test sets (Table 2). Both external validation criteria Q^2_{F1} and CCC were greater than the acceptance values (0.7 and 0.85, respectively) as defined by Chirico and Gramatica [43] indicating that the model may be accepted as externally predictive for new organic solvent. The fivefold cross-validation method used for internal validation purpose led to a satisfactory determination coefficient R^2_{CV} =0.729.

The best QSAR for algae pEC50 prediction was the following:

$$\begin{aligned} &Log\ 72\text{–}h\ algae\ EC50\ (mmol/L) \\ &\quad = 1.348(\pm 0.139) - 0.621(\pm 0.059)\ LogP \\ &\qquad + 0.249\ (\pm 0.095)\ LUMO \end{aligned} \tag{8}$$

Figure 4b shows the predicted vs experimental algae pEC50 both for training and test sets.

As expected and according to many authors [11, 13, 18, 20, 21, 73–76], the negative coefficient of LogP in the QSAR indicates that higher the lipophilicity, higher the toxicity of organic solvents toward algae. The positive coefficient of LUMO energy suggests that highly electrophilic compounds resulted in high toxicity in agreement with many authors and as observed in the QSAR developed for solvent fish acute toxicity solvents [26].

The Williams plot of the algae QSAR is presented in Fig. 5b. As for the developed QSAR for invertebrate pEC50 prediction, it can be clearly seen that the solvents are following a well-defined domain of applicability (with h* = 0.12). Only d-limonene was out of the AD due to its very high LogP value of 4.83. However, the corresponding predicted activity remains satisfactory with a corresponding standardized residual below the value 3. Once again, the predicted activity of the test solvents may be considered reliable.

For comparison, the ECOSAR relation (Eq. 6) was used to predict algae pEC50 of the 67 neutral organics solvents included in our dataset. Although Eq. 6 is based on activities measured over a test period of 96 h instead of 72 h for Eq. 8, Eq. 6 led to 0.640 and 0.571 for MAE and R^2 respectively which is in agreement with the determination coefficient $R^2 = 0.596$ characterizing Eq. 6 (see ECOSAR section). While the developed QSAR (Eq. 8) allowed reaching 0.766 for R^2 and 0.459 for MAE. Moreover, Eq. 8 remained usable to predict Algae pEC50 of the four esters remaining in the dataset with MAE = 0.492 against 0.886 for ECOSAR model (Eq. 6).

Conclusions

The prediction of the ecotoxicity profile of organic solvents by QSARs is of major relevance, especially to limit *in vivo* experiments. The description of aquatic toxicity requires the knowledge of the effects of a substance on organisms living in the water and represented by three trophic levels, i.e., vertebrates (fish), invertebrates (crustaceans as Daphnia spp.), and algae. In a previous work, we developed from a large dataset a reliable 4-parameter QSAR able to predict the fish LC50 of organic solvents, regardless of the mechanism of toxic action involved. Here, to complete this study and well-describe the ecotoxicity profile of organic solvents required by REACH regulation, we used the same approach to develop QSARs for the EC50 prediction of two other trophic levels, namely invertebrates and algae.

From the experimental activity data found in the literature and according to other studies, organic solvents showed similar toxicity toward fish and invertebrates while algae exhibited a different behavior. The 4-parameter QSAR developed for invertebrate pEC50 prediction included LogP, surface tension, dielectric constant, and the minimal atomic charge. As expected, this model is very similar to the one developed for fish LC50 prediction which includes LUMO energy instead of the minimal atomic charge. A 2-parameter QSAR involving LogP and LUMO energy allowed well-predicting algae pEC50 for all solvents other than amines which are well-known to exhibit specific toxicity behavior toward algae [68, 69]. These models have been obtained from a large solvent dataset and were validated by both internal and external validation techniques as required by the REACH regulation and according to OECD guidelines [52, 53]. They constitute a major tool for a reliable assessment of environmental risk related to organic solvents.

Acknowledgments This study was supported by the National Research Agency (ANR) within the project NESOREACH. The authors also gratefully acknowledge L. Geoffroy, L. Chancerelle and P. Pandard from the INERIS Institut.

References

1. Regulation No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH). Official J. of the European Union, L396/1-849, European Commission, Brussels, Belgium
2. Technical Guidance Document (TGD) on the Risk Assessment in support of the Commission directive 93/67/EEC on Risk Assessment for New Notified Substances, the Commission Regulation No 1488/94 on risk Assessment for Existing Substances and the Directive 98/8/EC concerning the placing of biocidal products on the market. Institute for health and consumer protection, European chemicals Bureau, Luxembourg
3. Working document. Guidance Document on Aquatic Ecotoxicology in the context of the Directive 91/414/ECC. European Commission, Health & Consumer protection directorate-generale
4. Katritzky AR, Lobanov VS, Karelson M (1995) QSPR: the correlation and quantitative prediction of chemical and physical properties from structure. Chem Soc Rev 24:279–287
5. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Mol Inf 29:476–488
6. Konemann H (1981) Quantitative structure-activity relationships in fish toxicity studies. Part 1: relationship for 5 industrial pollutants. Toxicology 19:209–221
7. Mazzatorta P, Vračko M, Jezierska A, Benfenati E (2003) Modeling toxicity by using supervised Kohonen neural networks. J Chem Inf Comput Sci 43:485–492
8. Katritzky AR, Slavovn SH, Stoyanova-Slavova S, Kahn I, Karelson M (2009) Quantitative structure-activity relationships (QSAR) modeling of EC50 of aquatic toxicities for Daphnia magna. J Toxicol Environ Health Part A 72:1181–1190
9. Toropova AP, Toropov AA, Benfenati E, Gini G (2012) QSAR models for toxicity of organic substances to Daphnia magna built up by using the CORAL Freeware. Chem Biol Drug Des 79:332–338
10. Hsieh SH, Hsu CH, Tsai DY, Chen CY (2006) Quantitative structure-activity relationships for toxicity of nonpolar narcotic chemicals to Pseudokirchneriella subcapitata. Environ Toxicol Chem 25:2920–2926
11. Lu G, Wang C, Tang Z, Guo X (2007) Joint toxicity of aromatic compounds to algae and QSAR study. Ecotoxicology 16:485–490
12. Kar S, Roy K (2010) QSAR modeling of toxicity of diverse organic chemicals to Daphnia magna using 2D and 3D descriptors. J Hazard Mater 177:834–840
13. Lee PY, Chen CY (2009) Toxicity and quantitative structure-activity relationships of benzoic acids to Pseudokirchneriella sucapitata. J Hazard Mater 165:156–161
14. Faucon JC, Bureau R, Faisant J, Briens F, Rault S (2001) Prediction of the Daphnia acute toxicity from heterogeneous data. Chemosphere 44:407–422
15. Moosus M, Maran U (2011) Quantitative structure-activity relationships analysis of acute toxicity of diverse chemicals to Daphnia magna with whole molecule descriptors. SAR QSAR Environ Res 22:57–774
16. Kamaya Y, Fukaya Y, Suzuki K (2005) Acute toxicity of benzoic acids to the crustacean Daphnia magna. Chemosphere 59:255–261
17. Chen CY, Kuo KL, Fan JW (2012) Toxicity of propargylic alcohols on green alga Pseudokirchneriella subcapitata. J Mol Struct 14: 181–186
18. Zeng M, Lin Z, Yin D, Zhang Y, Kong D (2011) A K(ow)-based QSAR model for predicting toxicity of halogenated benzenes to all algae regardless of species. Bull Environ Contam Toxicol 86:565–570
19. Passino-Reader DR, Hickey JP, Ogilvie LM (1997) Toxicity to Daphnia pulex and QSAR predictions for polycyclic hydrocarbons representative of Great Lake contaminants. Bull Environ Contam Toxicol 59:834–840
20. Lee YG, Hwang SH, Kim SD (2006) Predicting the toxicity of substituted phenols to aquatic species and its changes in the stream and effluent waters. Arch Environ Contam Toxicol 50:213–219
21. Lessigiarska I, Worth AP, Sohull-Klüttgen B, Jeram S, Dearden JC, Netzeva TI, Cronin MTP (2004) QSAR investigation of a large data set for fish, algae, and daphnia toxicity. SAR QSAR Environ Res 15:413–431
22. Voutchkova VM, Kostal J, Steinfeld JB, Emerson JW, Brooks BW, Anastas P, Zimmerman JB (2011) Towards rational molecular design: derivation of property guidelines for reduced acute aquatic toxicity. Green Chem 13:2373–2379
23. Niculescu SP, Lewis MA, Tigner J (2008) Probabilistic neural networks modeling of the 48-h LC50 acute toxicity endpoint to Daphnia magna. SAR QSAR Environ Res 19:735–750
24. Tao S, Xiaohuan X, Fuliu X, Bengang L, Cao J, Dawson R (2002) A fragment constant QSAR model for evaluating the EC50 values of organic chemicals to Daphnia magna. Environ Sci Technol 45: 4616–4622
25. Toropova AP, Toropov AA, Martyanov SE, Benfenati E, Gini G, Leszczynska D, Leszczynska J (2012) CORAL: QSAR modeling of toxicity of organic chemicals toward Daphnia magna. Intell Lab Syst 110:177–181
26. Levet A, Bordes C, Clément Y, Mignon P, Chermette H, Marote P, Cren-Olivé C, Lantéri P (2013) Quantitative structure-activity relationship to predict acute fish toxicity of organic solvents. Chemosphere 93:1094–1103
27. Chastrette M, Rajzmann M, Chanon M, Purcell KF (1985) Approach to a general classification of solvents using a multivariate statistical treatment of quantification solvent parameters. J Am Chem Soc 107:1–11
28. Durand M, Molinier V, Kunz W, Aubry JM (2011) Classification of organic solvents revisited by using the COSMO-RS approach. Chem Eur J 17:5155–5164
29. Cheremisinoff NP (2003) Industrial solvents handbooks, second edn. CRC, Boca Raton
30. Smallwood I (1996) Handbook of organic solvent properties. Arnold
31. Yalkowxky SH, He Y, Jain P (2010) Handbook of aqueous solubility data, second edn. CRC, Boca Raton

32. Yaws CL (1999) Chemical properties handbook. McGraw-Hill, New York
33. Weast RC (1975) Handbooks of chemistry and physics, 56th edn. CRC, Boca Raton
34. Perdew JP, Burke K, Ernzerhof M (1996) Generalized gradient approximation made simple. Phys Rev Lett 77:3865–3868
35. Wiener H (1947) Structural determination of paraffin boiling points. J Am Chem Soc 69:17–20
36. Balaban AT (1982) Highly discriminating distance-based topological index. Chem Phys Lett 89:399–404
37. Randić M (1975) On characterizaztion of molecular branching. J Am Chem Soc 97:6609–6611
38. Kier LB (1989) An index of molecular flexibility from kappa shape attributes. Quant Struct –Act Relat 8:221–224
39. Mercader AG, Duchowicz PR, Fernandez FM, Castro EA (2008) Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories. Chemom Intell Lab Syst 92:138–144
40. Kubinyi H (1994) Variable selection in QSAR studies. I. An evolutionary algorithm. Quant Struct-Act Relat 13:285–294
41. Johnson ME, Moore LM, Ylvisaker D (1990) Minimax and maximim distance designs. J Stat Plan Int 26:131–148
42. Chirico N, Gramatica P (2011) Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J Chem Inf Model 51:2320–2335
43. Chirico N, Gramatica P (2012) Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection. J Chem Inf Model 52:2044–2058
44. Shi LM, Fang H, Tong W, Wu J, Perkins R, Blair RM, Branham WS, Dial SL, Moland CL, Sheehan DM (2001) QSAR models using a large diverse set of estrogens. J Chem Inf Comput Sci 41: 186–195
45. Schüürmann G, Ebert RU, Chen J, Wang B, Kühne R (2008) External validation and prediction employing the predictive squared correlation coefficient—Test set activity mean vs training set activity mean. J Chem Inf Model 48:2140–2145
46. Consonni V, Ballabio D, Todeschini R (2009) Comments on the definition of the Q^2 parameter for QSAR Validation. J Chem Inf Model 49:1669–1678
47. Consonni V, Ballabio D, Todeschini R (2010) Evaluation of model predicitive ability by external validation techniques. J Chemometrics 24:194–201
48. Roy PP, Paul S, Mitra I, Roy K (2009) On-two novel parameters for validation of predictive QSAR models. Molecules 14:1660–1701
49. Roy K, Mitra I, Kar S, Ojha PK, Das RN, Kabir H (2012) Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model 52:396–408
50. Golbraikh A, Tropsha A (2002) Beware of q^2! J Mol Graph Model 20:269–276
51. Lin LI (1989) A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268
52. OECD (2007) Series on testing and assessment. Number 69. Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models
53. OECD (2004) Series on testing and assessment. Number 49. The report from the expert group on QSARs on the principles for the validation of QSARs
54. Eriksson L, Jaworska J, Worth AP, Cronin MTD, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for Applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375
55. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701
56. Voutchkova AM, Osimitz TG, Anastas PT (2010) Toward a comprehensive molecular design framework for reduced hazard. Chem Rev 110:5845–5882
57. 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
58. Reuschenbach P, Silvani M, Dammann M, Warnecke D, Knacker T (2008) ECOSAR model performance with a large test set of industrial chemicals. Chemosphere 71:1986–1995
59. Tunkel J, Mayo K, Austin C, Hickerson A, Howard P (2005) Practical considerations on the use of predictive models for regulatory purposes. Environ Sci Technol 39:2188–2199
60. Parkerton TF, Konkel WJ (2000) Application of quantitative structure-activity relationships for assessing the aquatic toxicity of phthalate esters. Ecotoxicol Environ Saf 45:61–78
61. Hodges G, Roberts DW, Marshall SJ, Dearden JC (2006) The aquatic toxicity of anionic surfactants to Daphnia magna—a comparative QSAR study of linear alkylbenzene sulphonates and ester sulphonates. Chemosphere 63:1443–1450
62. Buth JM, Arnold WA, McNeill K (2007) Unexpected products and reaction mechanisms of the aqueous chlorination of cimetidine. Environ Sci Technol 41:6228–6233
63. Tremolada P, Finizio A, Villa S, Gaggi C, Vighi M (2004) Quantitative inter-specific chemical activity relationships of pesticides in the aquatic environment. Aquat Toxicol 67:87–103
64. Zhang XJ, Qin HW, Su LM, Qin WC, Zou MY, Sheng LX, Zhao YH, Abraham MH (2010) Interspecies correlations of toxicity to eight aquatic organisms: theoretical considerations. Sci Total Environ 408:4549–4555
65. Netzeva TI, Aptula AO, Benfenati E, Cronin MTD, Gini G, Lessaigiarska I, Maran U, Vračko M, Schüürmann G (2005) Description of the electronic structure of organic chemicals using semiempirical and ab initio methods for development of toxicological QSARs. J Chem Inf Model 45:106–114
66. Henegar A, Mombelli E, Pandard P, Péry ARR (2011) What can be learnt from an ecotoxicity database in the framework of the REACh regulation? Sci Total Environ 409:489–494
67. Tebby C, Mombelli E, Pandard P, Péry ARR (2011) Exploring an ecotoxicity database with the OCDE (Q)SAR Toolbox and DRAGON descriptors in order to priorise testing on algae, daphids and fish. Sci Total Environ 409:3334–3343
68. Christensen AM, Faaborg-Andersen S, Ingerslev F, Baun A (2007) Mixture and single- substance toxicity of selective serotonin reuptake inhibitors toward algae and crustaceans. Environ Toxicol Chem 26:85–91
69. Escher BI, Bramaz N, Richter I, Lienert J (2006) Comparative ecotoxicological hazard assessment of Beta-flockers and their human metabolites using a mode-of-action based test battery and a QSAR approach. Environ Sci Technol 40:7402–7408
70. Von der Ohe PC, Kühne R, Ebert R-U, Alterburger R, Liess M, Schüürmann G (2005) Structural alerts- a new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnids assay. Chem Res Toxicol 18:535–555
71. Papa E, Villa F, Gramatica P (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow). J Chem Inf Model 45:1256–1566
72. Neuwoehner J, Escher BI (2011) The pH-dependent toxicity of basic pharmaceuticals in the green algae Scendesmus vacuolatus can be explained with a toxicokinetic ion- trapping model. Aquat Toxicol 101:266–275
73. Huang CP, Wang Y-J, Chen C-Y (2007) Toxicity and quantitative structure activity relationships of nitriles based on Pseudokirchneriella sucapitata. Ecotoxicol Environ Saf 67:439–446

74. Lu G-H, Yuan X, Zhao Y-H (2001) QSAR study on the toxicity of substituted benzenes to the algae (Scenedesmus obliquus). Chemosphere 44:437–440

75. Schmitt H, Alterburger R, Jastorff B, Schüürmann G (2000) Quantitative Structure-activity analysis of the algae toxicity of nitroaromatic compounds. Chem Res Toxicol 13:441–450

76. Zhu M, Ge F, Zhu R, Wang X, Zheng X (2010) A DFT based QSAR study of the toxicity of quaternary ammonimum compounds on Chlorella vulgaris. Chemosphere 80:46–52