CASE STUDY

Development and Validation of QSAR Models for (B)TAZs and Fragrances

Simona Kovarich

QSAR Research Unit
in Environmental Chemistry and Ecotoxicology
www.qsar.it
University of Insubria - Varese

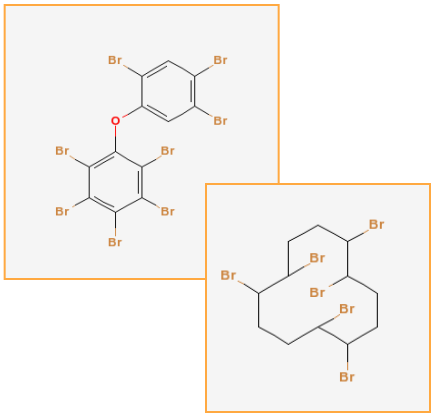
e-mail: simona.kovarich@uninsubria.it

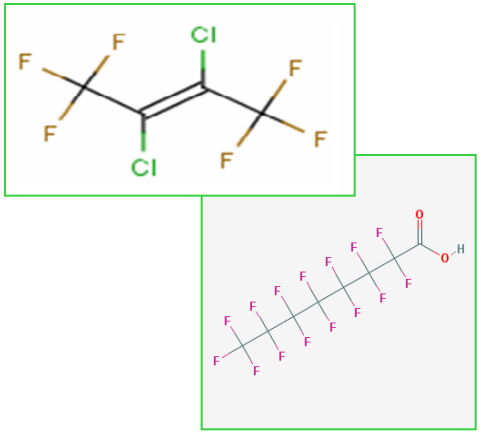
CADASTER Workshop on the development and application of QSAR models in REACH
7-9 October 2012, Munich

EU-FP-7 CADASTER

Focus on 4 classes of emerging pollutants

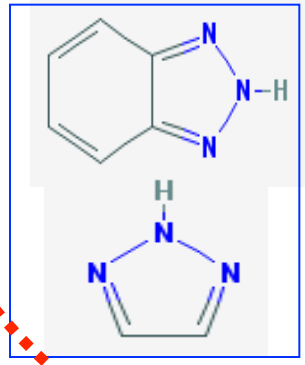
Brominated flame retardants (PBDEs, TBBPA, etc..)

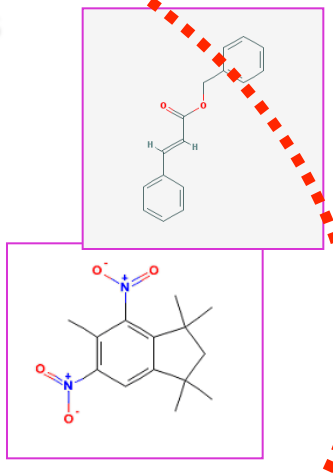




Poly- and Perfluorinated compounds (PFCs)

(Benzo-)Triazoles (B)TAZs





Substituted musks/ Fragrances

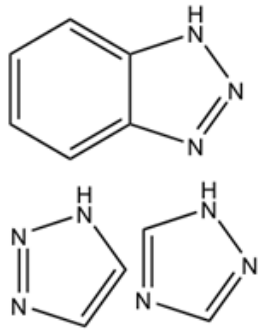
→ **QSAR models were developd (WP3) according to the available experimental data (limited)**

- ❖ **Prioritization for experimental test (WP2)**
- ❖ **Integration of QSAR predictions for risk assessment (WP4)**

CASE STUDY 1

QSAR models for aquatic toxicity of Triazoles and Benzo-triazoles (B)TAZs

Triazoles and Benzo-triazoles (B)TAZs



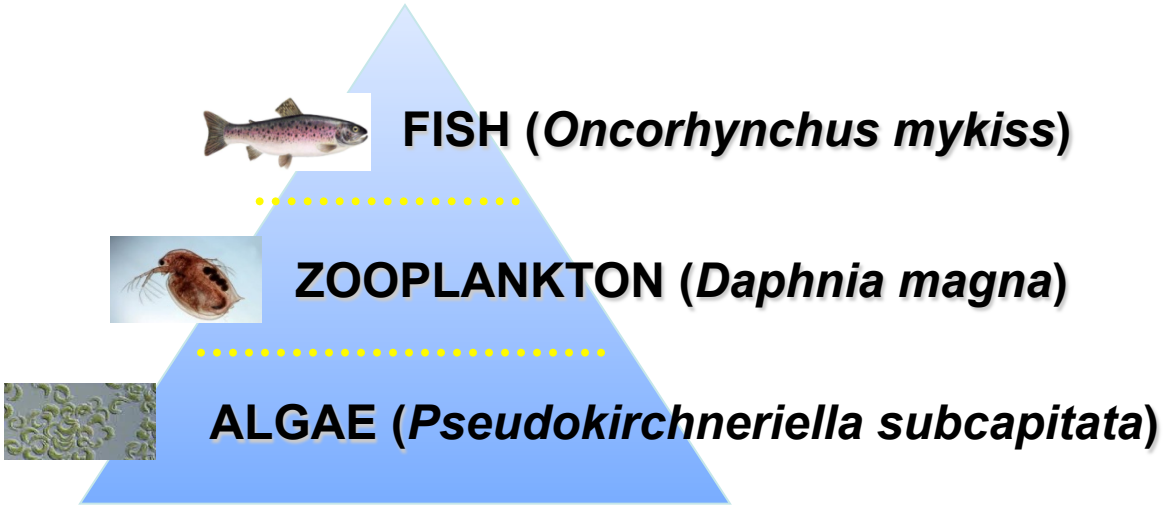
- Synthetic compounds, structurally heterogeneous
- Presence of an aromatic heterocycle (2 C + 3N atoms)

- Wide applications and use: phytosanitary products, pharmaceuticals, cleaning agents for textiles, UV stabilizer for plastics, de-icing agents, etc.

- High stability and environmental persistance
- High water solubility (contamination of water compartment)

- Experimental data for the basic ecotoxicological endpoint available for several (B)TAZs used as pesticides.
- Focus on toxicity of (B)TAZs in the aquatic environment.

- Development of QSAR models for the predicion of acute toxicity of (B) TAZs to aquatic organisms



Linnæus University Sweden

Swedish Environmental Research Institute

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

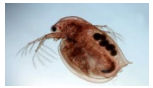
(B)TAZs - Materials and methods

DATA SET

- Experimental data collected from the Footprint database (PPDB)
- Data collected for (B)TAZs and other azo-aromatic compounds (e.g. diazines and triazines)

ENDPOINTS

ALGAE: *Pseudokirchneriella subcapitata*
log 1/EC_{50} 72 h (growth inhibition) **N = 35**



ZOOPLANKTON: *Daphnia magna*
log 1/EC_{50} 48 h (immobilization) **N = 97**

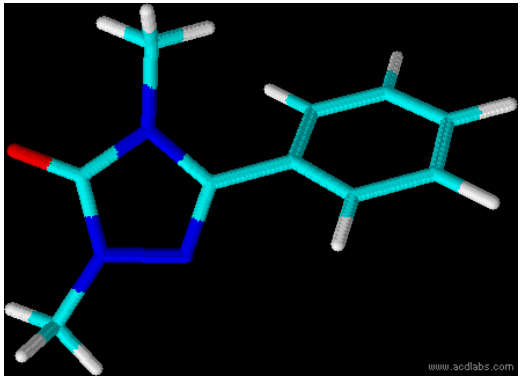
FISH: *Oncorhynchus mykiss*
log 1/LC_{50} 96 h **N = 75**

	UI	IVL	LnU	HMGU	IDEA
Endpoint	EC_{50} algae EC_{50} *Daphnia* LC_{50} fish	EC_{50} algae EC_{50} *Daphnia* LC_{50} fish	EC_{50} algae EC_{50} *Daphnia* LC_{50} fish	EC_{50} algae EC_{50} *Daphnia* LC_{50} fish	EC_{50} *Daphnia* LC_{50} fish
Algorithm	MLR-OLS	PLSR	PLSR and BLASSO-PLS	kNN, ASNN, FSMLR, PLS, MLRA, SVM	MLR-OLS
Molecular Descriptors	Dragon 5.5, PaDEL, CADASTER (1D-2D)	Dragon 6 (1D-2D-3D)	Dragon 6 (1D-2D-3D)	CADASTER (1D-2D-3D)	Dragon 5.5 (1D-2D)
Applicability Domain	Leverage	DModX	Leverage on PLS latent variables	STD of ASNN	Leverage
Validation	Internal (R^2, Q^2_{LOO}, Y-sc etc) External (Q^2_{ext} F1-F2-F3, CCC etc)	Internal (R^2,Q^2, RMSEE) External (RMSEP)	Internal (R^2, RMSE) External (Q^2_{ext}, RMSE)	Internal and External (R^2, Q^2, RMSE and MAE)	Internal (R^2, Q^2_{LOO}, Y-sc etc) External (Q^2_{ext} F1-F2-F3, CCC)

UI approach

Structures and Descriptors

- Molecular structures were drawn and minimized by the semi-empirical method AM1 (HYPERCHEM software), and converted into SMILES and MOL (Open Babel)



- Mono- and bi-dimensional descriptors were calculated using Dragon (v. 5.5), PADEL-Descriptor (v. 2.13) and QSPR-Thesaurus (CADASTER on-line platform)

CAS	AMW	Sv	Ss	Mv	Me	Ms
000050-29-3	12.66	21.69	45.81	0.77	1.03	2.41
000050-30-6	12.73	11.22	33.89	0.75	1.06	3.08
000050-31-7	15.03	11.92	37.67	0.79	1.09	3.14
000050-32-8	7.89	23.59	37.33	0.74	0.98	1.87
000051-28-5	10.83	11.14	49	0.66	1.1	3.77
000051-44-5	12.73	11.22	33.89	0.75	1.06	3.08
000055-38-9	8.98	19.38	35.81	0.63	1.01	2.24
000055-63-0	10.77	11.67	60.83	0.56	1.14	4.06
000056-23-5	30.76	5	17.69	1	1.21	3.54
000056-38-2	9.1	19.71	49.31	0.62	1.03	2.74
000057-15-8	11.83	9.6	24.83	0.64	1.05	3.1
000057-74-9	17.07	19.79	46.81	0.82	1.07	2.6
000058-89-9	16.16	13.79	32.67	0.77	1.07	2.72
000058-90-2	17.84	11.11	32.78	0.85	1.1	2.98
000059-50-7	8.91	10.6	23.11	0.66	1.01	2.57
000060-29-7	4.94	7.5	10.5	0.5	0.98	2.1
000060-51-5	9.55	14.17	31.97	0.59	1.02	2.66

Algorithm

- Multiple linear regression (MLR) performed by Ordinary Least Squares (OLS) method.
- Variable selection by Genetic Algorithm (GA).

Qi
QSARINS

Tools of Validation

- **Internal stability verified by R^2, Q^2_{LOO}, Q^2_{LMO}, R^2/Q^2_{YS}, etc.**
- **External predictivity verified by different Q^2_{EXT} criteria [Q^2_{EXT} F1-F2-F3 and CCC]**

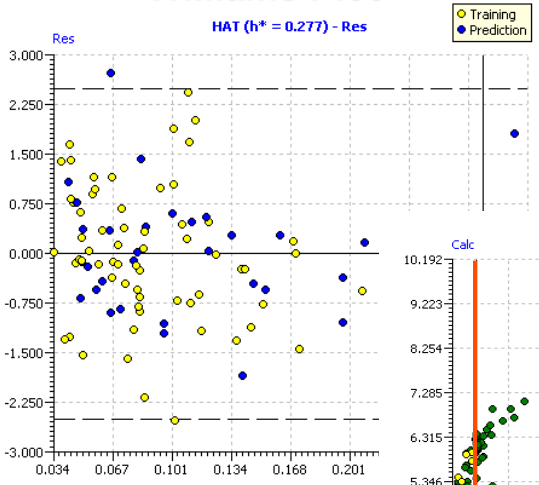
Qi
QSARINS

Prediction sets were obtained by splitting 30% (K-ANN, random by response). An additional external validation set (EV set) was also used in the fish model (18 (B)TAZs).

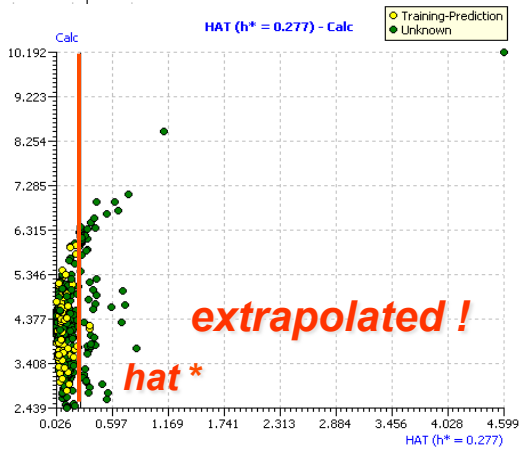
Applicability Domain

- **Structural applicability domain verified by the Leverage approach**
- **Interpolated and extrapolated predictions verified by Leverage**

Williams Plot



Insubria Graph

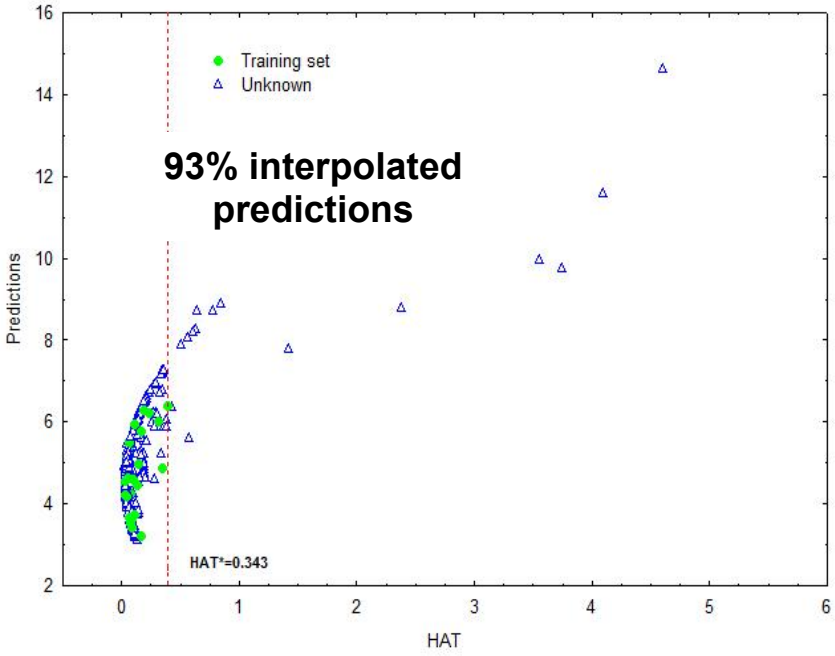


QSARs for *P. subcapitata* EC_{50} 72h

Model	Descriptors	R^2	Q^2_{LOO}	Q^2_{LMO}	Q^2_{EXT-Fn}	CCC	SPLIT MODELS
Dragon	*AEigZ, T(N..S), SEigv*	82%	77%	77%	72-84%	86-87%	
QSPR-Thesaurus	*p1p4-5N, C-C, p5BE*	80%	73%	74%	71-83%	86-87%	
PaDEL-Descriptor	*AMR, MDEN-22, maxHBa*	82%	76%	76%	67-89%	82-91%	

- Starting from different pool of descriptors, the GA selected molecular descriptors encoding for similar structural information.

- Screening of 369 (B)TAZs without experimental data (ECHA pre-registration list)



Publication in press in *Molecular Informatics* ...

Importance of data curation

Publication in press in *Molecular Informatics*

Experimental data quality and variability

Y = *f* (X)

Chemical structures correctness and Input formats for the calculation of descriptors

pEC_{50}	PPDB	WP2 (PHI)	ΔpEC_{50}
Triadimefon	5.16	4.59	0.57
Epoxiconazole	5.44	4.58	0.86

- Despite both high quality data and same method used (OECD guidelines), variablity in experimental data can affects the QSAR model.

Specific attention to experimental uncertainty, input and basis of QSAR models, is therefore necessary

- Differences in chemical structures generated from SMILES collected from different sources (Pubchem, ChemID plus, Chemspider, etc...)

- Differences in molecular descriptors (nH, nAB, nitro groups) calculated using different software and different input files (SMILES, MOL, HIN)

Accurate check of the structures and harmonize the representation of specific groups

QSARs for *D. magna* EC_{50} 48h

Model	Descriptors	N	R^2	Q^2_{LOO}	Q^2_{EXT-Fn}	CCC	
Dragon	*TPSA(NO), Aeigm, nCar, nHDon, H-052*	97	77%	74%	69-83%	85-89%	SPLIT MODELS
QSPR-Thesaurus	*ALogP p5-3N C(-H'-C'=C')*	97	75%	73%	71-78%	86-87%	

QSARs for *O. mykiss* LC_{50} 96h

Model	Descriptors	N	R^2	Q^2_{LOO}	Q^2_{EXT-Fn}	CCC	
Dragon	*CIC1, Mp, H-052, TPSA (tot)*	75	79%	76%	85%	92%	EV set
PaDEL-Descriptor	*VP-1, SHBint2, maxHaaCH*	75	76%	73%	71-72%	82%	

Models applied to predict acute toxicity of > 300 (B)TAZs without experimental data (> 90% included in AD of models)

WP3 Consensus predictions

❖ **Comparison of individual WP3 Models**

comparable predictions for (B)TAZs included in the AD of all the models.

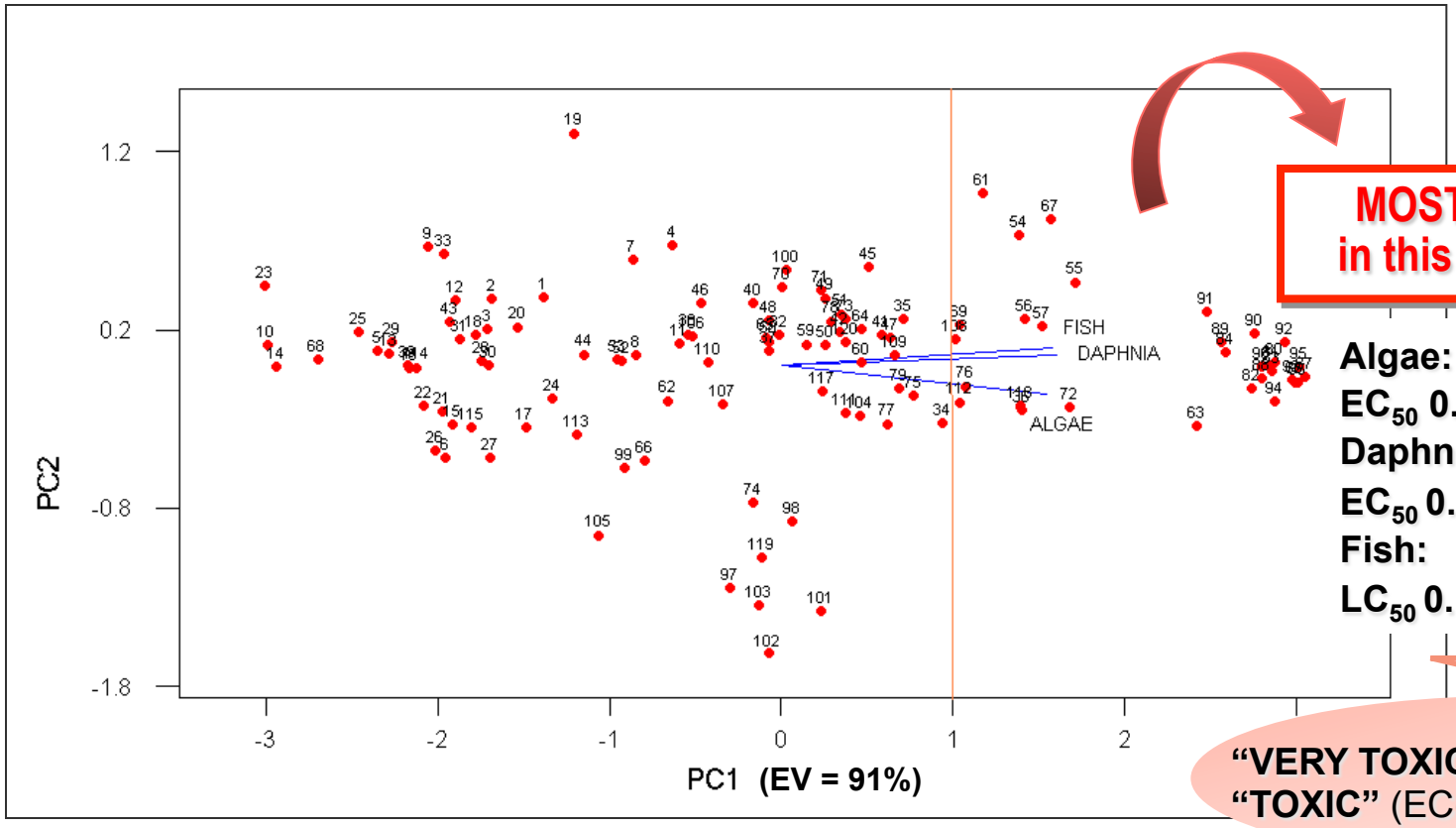
❖ **WP3 Consensus Models**

combination of predictions for 386 (B)TAZs obtained from different WP3 models and approaches, taking into account statistical performances and applicability domains of individual models.

❖ **Screening of (B)TAZs**

consensus predictions (algae, daphnia, fish) analysed by PCA in order to screen the studied (B)TAZs according to their toxicological profile in the aquatic environment.

Screening and prioritization of (B)TAZs



MOST ACTIVE BTAZs in this aquatic scenario

Algae:
EC_{50} 0.18 – 4.62 mg/L
Daphnia:
EC_{50} 0.19 – 5.37 mg/L
Fish:
LC_{50} 0.12 – 4.65 mg/L

"VERY TOXIC" ($EC(LC)_{50} \leq 1$ mg/L)
"TOXIC" ($EC(LC)_{50} \leq 10$ mg/L

TREND of AQUATIC TOXICITY

(B)TAZs - Conclusions

- **Robust and externally predictive QSAR models have been developed to predict the acute toxicity of (B)TAZs in Algae, *Daphnia* and Fish, using both commercial and freely available molecular descriptors.**

- **Models were applied to predict aquatic toxicity of > 350 (B)TAZs without experimental data (ECHA list): interpolated predictions for 90% of compounds.**

- **Importance of data curation for the development of valid QSARs**

- **Importance of the consensus approach to increase reliability of QSAR predictions.**

APPLICABILITY OF THE PROPOSED MODELS…

Acute toxicity on algae, daphnia and fish are the basic endpoints required in REACH for risk assessment of chemicals (→ prediction used in WP4)

Models will be implemented in the CADASTER database: they will be freely available for users to be applied also for regulatory purposes

CASE STUDY 2

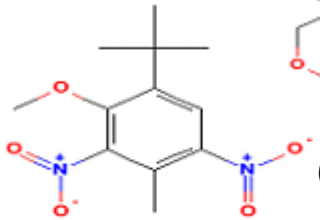
QSAR models for biodegradation of Fragrances

Fragrances

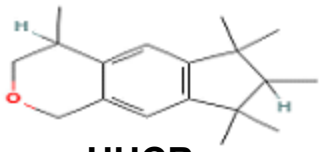
- Compounds widely used in cosmetics, toiletries, household and loundry products, products used to scent the air (air freshners and fragranced candles), food additives.

- Structurally highly heterogeneous (22 major structural classes identified, exhibiting a wide range of physico-chemical properties).

Musks

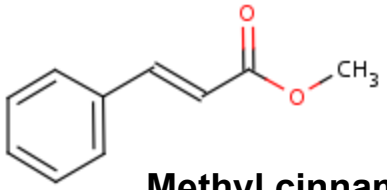


Musk ambrette



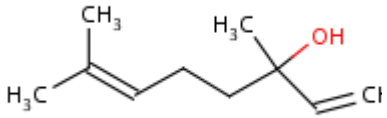
HHCB (Galaxolide)

Cinnamic Acid derivatives



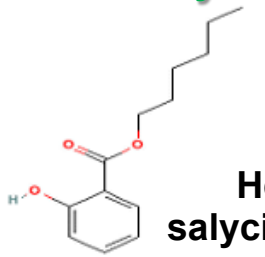
Methyl cinnamate

Terpens



Linalool

Salicylates



Hexyl salycilate

- Wide use and exposure, but limited information available related to health effects of fragrances. Fragrance formulas considered as trade secrets.

- Known effects: asthma, allergies and migraine headaches, accumulate in adipous tissues (breast milk), some fragrances suspected of being endocrine disruptors.

Biodegradation

❖ It's one of the main processes for removal of chemicals from the environment.

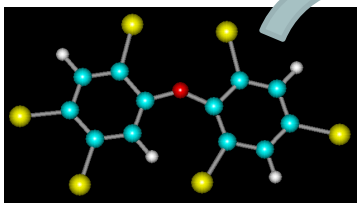
❖ The "benign by design" concept requires information on a compound's biodegradability to be available at an early stage, before synthesis of new chemicals.

❖ Use of *in-silico* techniques (eg. QSAR) for a rational design of new fragrances

❖ Classification models

$$C = f(X_i)$$

Quantitative relationship between structure (X_i) and a qualitative response (C)



Molecular descriptors

CAS	AMW	Sv	Ss
000050-29-3	12.66	21.69	45.81
000050-30-6	12.73	11.22	33.89
000050-31-7	15.03	11.92	37.67
000050-32-8	7.89	23.59	37.33
000051-28-5	10.83	11.14	49
000051-44-5	12.73	11.22	33.89
000055-38-9	8.98	19.38	35.81
000055-63-0	10.77	11.67	60.83
000056-23-5	30.76	5	17.69
000056-38-2	9.1	19.71	49.31

2 Classes:

- **Ready Biodegradable (RB)**
- **Not Ready Biodegradable (NRB)**

→ Classification method: *k Nearest Neighbour* (*k*-NN)

- searches for the k nearest neighbours of each molecule in the dataset
- compound assigned to the most representative class of the *k* neighbours (1<*k*<10)

→ Molecular descriptors: commercial software (Dragon, v. 5.5.) and *free on-line* (PaDEL-Descriptor 2.12)

❖ Model Performances and Validation

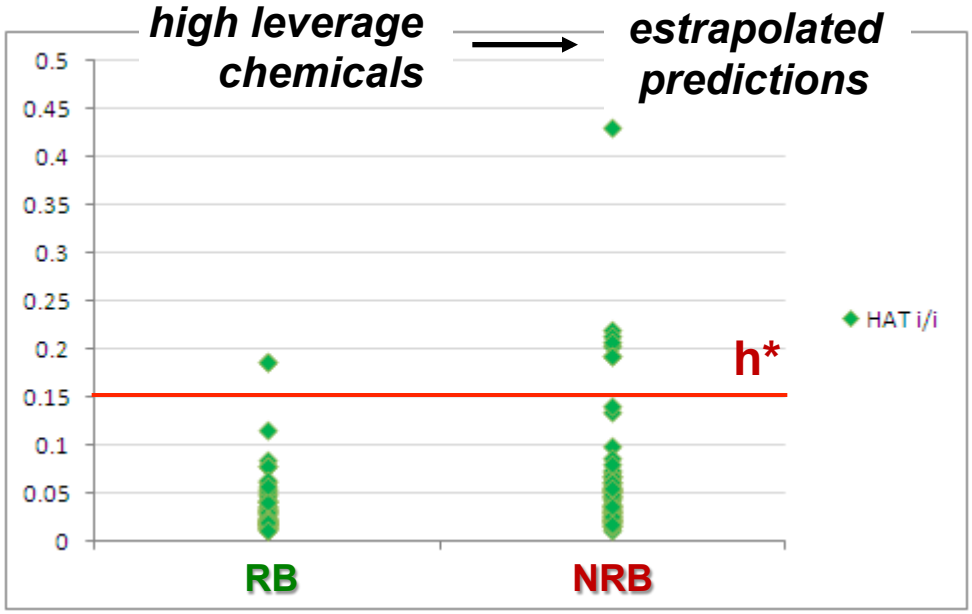
	Assigned Class		
Real Class	RB	NRB	**Accuracy**
RB	✓	✗	% RB
NRB	✗	✓	% NRB
			% Overall Accuracy (OA)

External validation → **predictivity**

❖ Applicability Domain

→ **applicability of the model to new molecules**

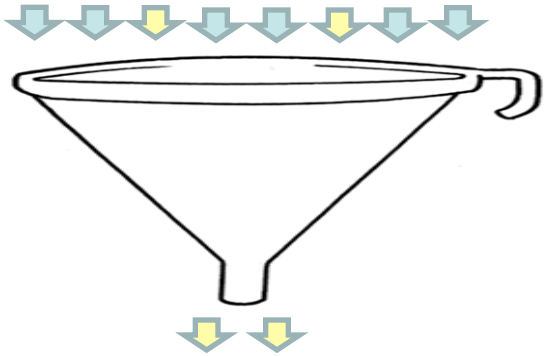
- **Range of descriptors**
- ***Leverage***

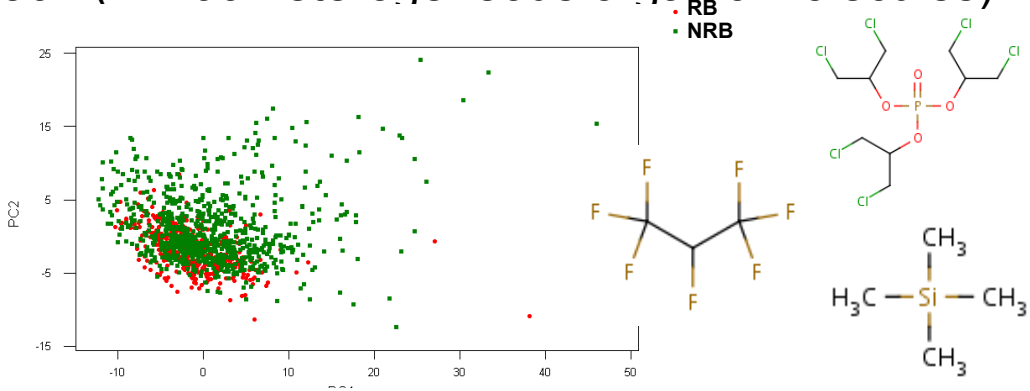


❖ DATA SET

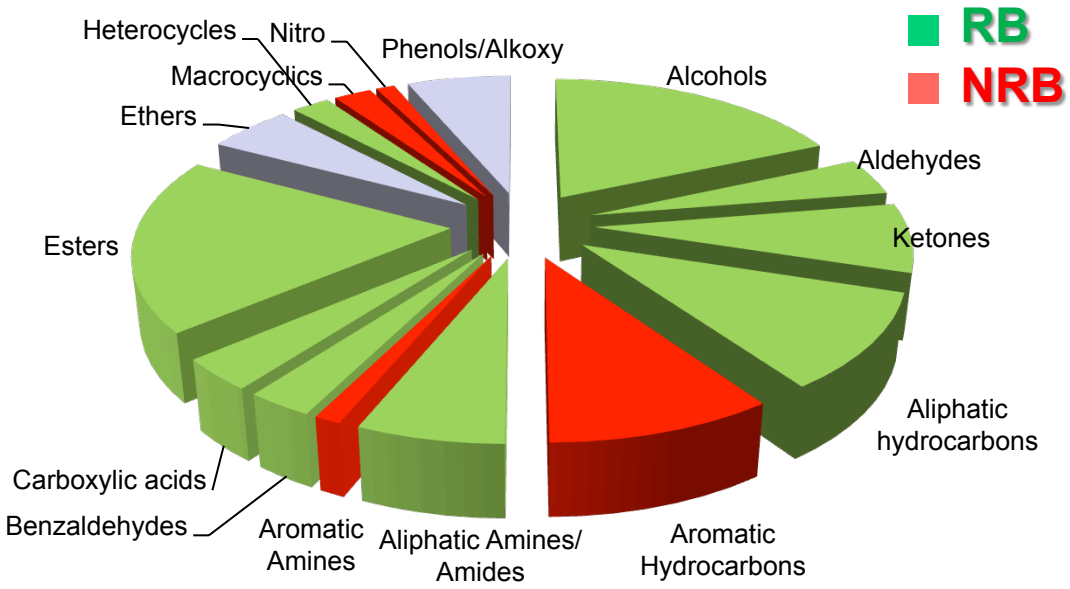
Japanese MITI (Ministry of International Trade and Industry) **database**

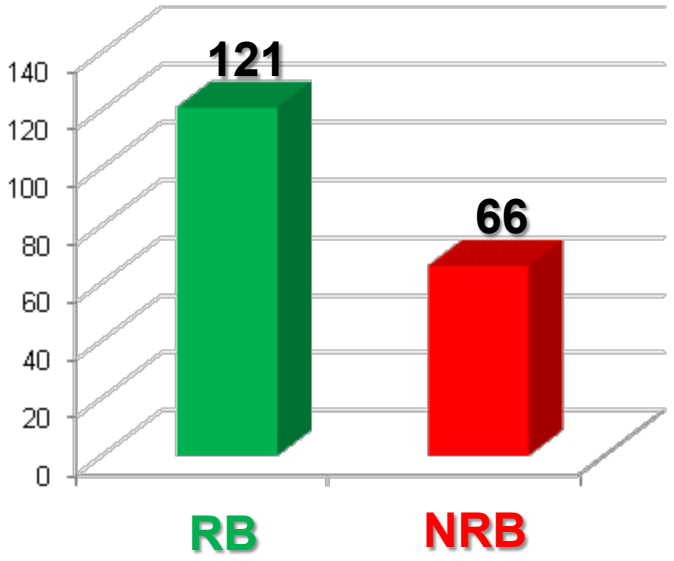
→ *ready biodegradability* data OECD 301 (~1400 heterogeneous organic molecules)



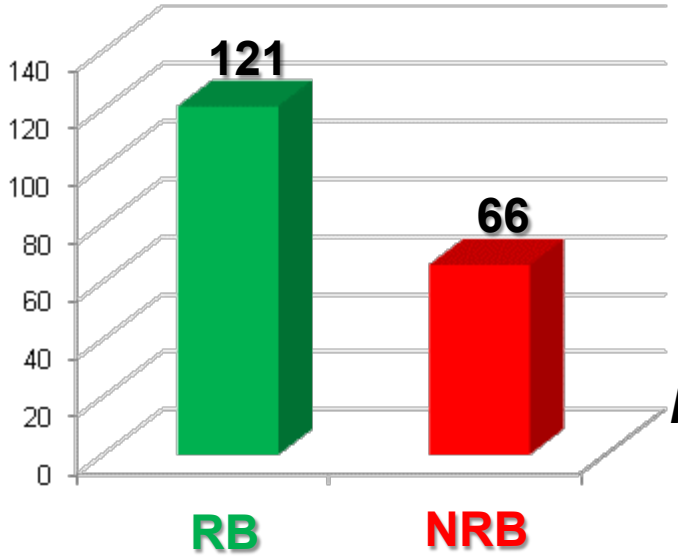


Data set: 187 compounds (100 fragrances)

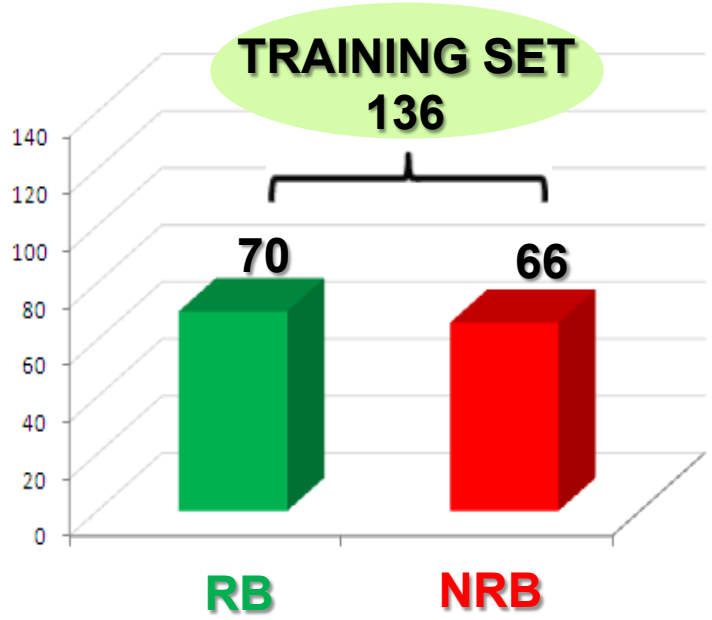




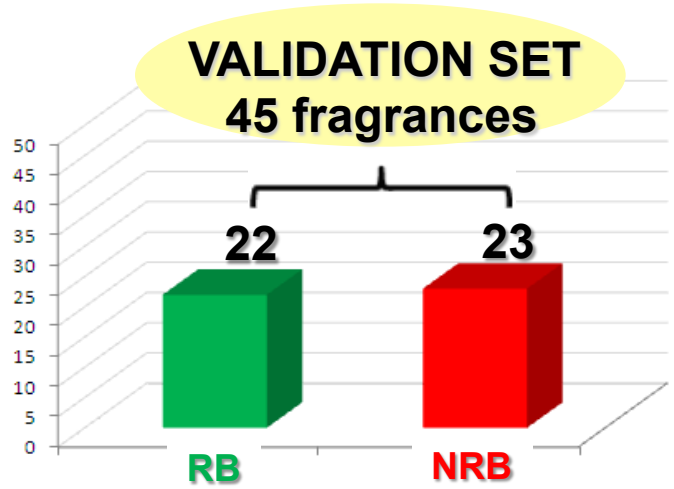
❖ DATA SET balancing



Factorial Analysis



❖ External validation



- **35 data from RIFM (*Research Institute for Fragrance Materials*)**
- **10 data measured within CADASTER - WP2 (PHI, *Public Health Institute Maribor*)**

❖ Best models

DRAGON Models

		Training set (n=136)			Validation set (n=45)		
Model ID	*k*	RB%	NRB%	OA%	RB%	NRB%	OA%
M1	5	85.7	72.7	79.4	72.7	60.9	66.67
M2	4	82.9	77.3	80.1	68.2	60.9	64.4
M3	6	88.6	78.8	83.8	72.7	69.6	71.1
Consensus		94.3	80.3	87.5	72.7	73.9	73.3

PaDEL-Descriptor Models

Model ID	*k*	RB%	NRB%	OA%	RB%	NRB%	OA%
M4	7	81.4	78.8	80.1	72.7	73.9	73.3

BioWIN (EPI Suite)

RB%	NRB%	OA%	RB%	NRB%	OA%
72.8	62.1	67.6	63.6	78.3	71.1

❖ Interpretation of modeling descriptors

DRAGON

PaDEL →

Model ID	Molecular Descriptors
M1	nClC, X0A, nR=Ct, F01(C-O)
M2	TI1, Vindex, nCq, H-052, B06(C-O)
M3	Sv, Qindex, MAXDP, GGI5, nCq
M4	maxHBa, maxHssNH, maxssssC, WTPT-2

RB

- Alcohols, aldeydes, ketones, esters ...
- amines

NRB

- n° rings (aromatic)
- n° tertiary/quaternary Carbons

FRAGRANCES - Conclusions

- **Robust and externally predictive classification QSARs have been developed for the prediction of ready biodegradability of fragrances.**

- **QSARs based on both commercial (Dragon) and freely available (PaDEL-Descriptor) software.**

- **Automatic selection of molecular descriptors (GA), starting from hundreds of descriptors, is able to select really relevant descriptors for biodegradation.**

APPLICABILITY OF THE PROPOSED MODELS…

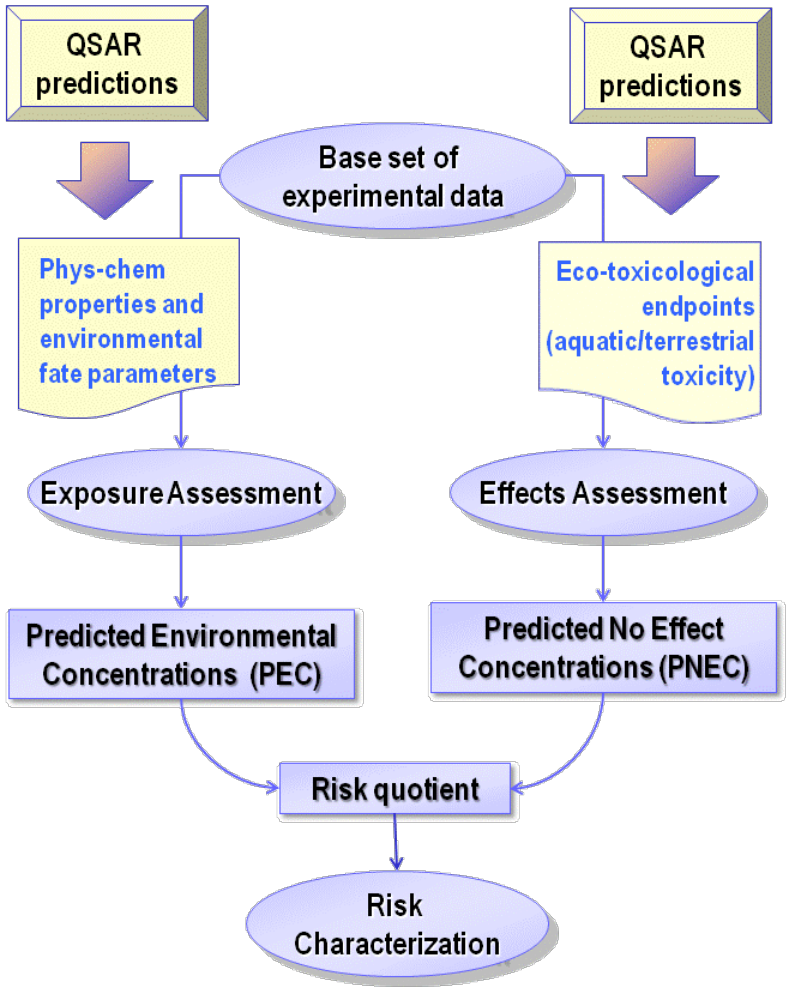
Ready biodegradation is among the basic information concerning the environmental fate of chemicals required for risk assessment.

Screening of numerous fragrances and *a priori* use in the design of new alternative compounds, which are less persistent according to the "green chemistry" philosophy

Novel and predictive QSAR models for specific classes of hazardous emerging pollutants with limited data availability.

Potential applications:

- ✓ filling data gaps
- ✓ evaluation of chemicals of interest for regulation.
- ✓ support tools for environmental risk assessment.
- ✓ screening pre-synthesis of chemicals

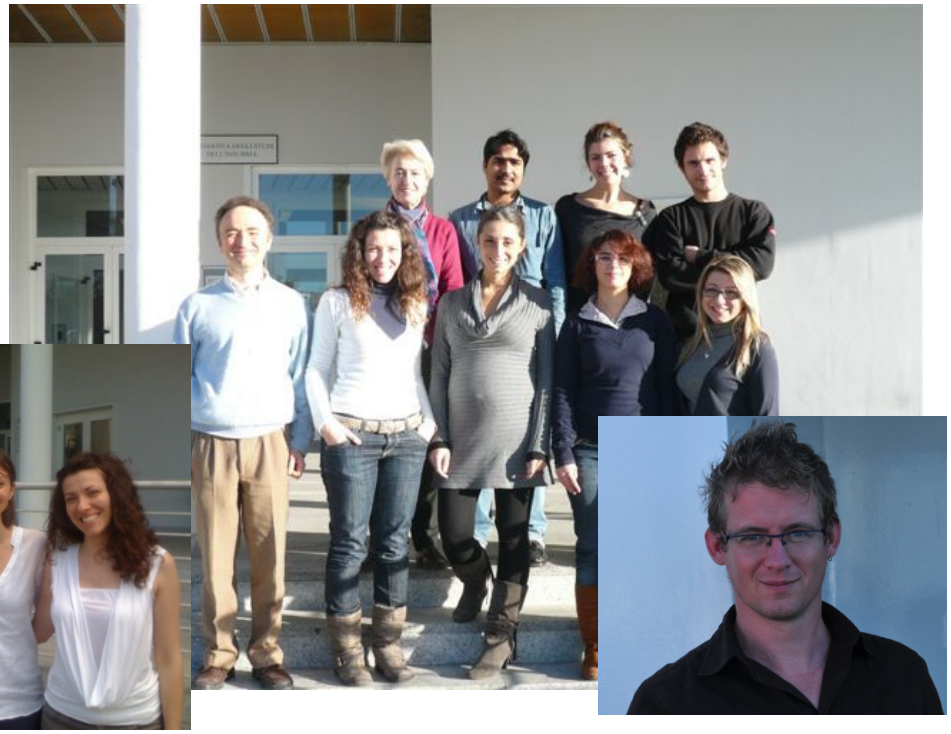


Acknowledgments

Financial support by the European Union through the project CADASTER (FP7-ENV-2007-212668).

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, University of Insubria, Varese (ITALY)

Prof. Paola Gramatica
Dr. Ester Papa
Dr. Leon van der Wal
Dr. Nicola Chirico
Stefano Cassani
Lidia Ceriani
Alessandra Rizza



Thank you for your attention!