QSAR prediction of aquatic and mammalian toxicity of triazoles and benzo-triazoles

Ester Papa[1], Stefano Cassani[1], Elisa D'Onofrio[1], Simona Kovarich[1], Partha Pratim Roy[1], Magnus Rahmberg[2], Sara Nilsson[2], Paola Gramatica[1]

[1] QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, Varese, Italy.; [2] IVL Swedish Environmental Research Institute Ltd.

e-mail: ester.papa@uninsubria.it; paola.gramatica@uninsubria.it

INTRODUCTION & OBJECTIVES

Triazoles and benzo-triazoles (BTAZs) are potentially hazardous chemicals that adversely affect humans and other non-target species, and are on the list of substances of very high concern (SVHC) in the European regulation of chemicals REACH.

TAZ/BTAZs are synthetic molecules used in various industrial processes (to obtain pharmaceuticals and agricultural products), and have a wide application as anti-corrosives, cleaning agents for textiles, flame retardants, photographic emulsions, etc... Furthermore they are abundantly used as components of liquid deicing agents for aircraft and airport runways. Because of their wide use they have been found distributed throughout the environment, mainly in water compartments. The amount of experimental data available for these molecules is insufficient for a comprehensive characterization of their environmental and toxicological profile and they have been included among the four classes of chemicals studied in the European FP7 Project CADASTER (CAse studies on the Development and Application of in Silico Techniques for Environmental hazard and Risk assessment) [1].

OBJECTIVES:

- Development of QSAR models, by different modeling approaches (MLR OLS and PLS), for the three key organisms for the aquatic ecosystem (Alga, *Daphnia* and Fish), in order to define the potential aquatic toxicological profile of BTAZs.
- Definition of *Daphnia*-Fish Interspecies Quantitative Correlation.
- Development of QSAR models for the mammalian toxicity of BTAZs.

MATERIALS & METHODS

	UI	IVL
DATA SETS	Experimental data of aquatic and mammalian toxicity of BTAZs (also included in the ECHA list) were collected from the Footprint database [2]. Additional data for azo-aromatic compounds (diazines, triazines and similar heterocycles) were collected to improve the robustness and the predictivity of UI models for Daphnia and Fish.	
ENDPOINTS	Algae (*Pseudokirchneriella subcapitata*): EC50 96h; Daphnia (*Daphnia magna*): EC50 48h; Fish (*Oncorhynchus mykiss*): LC50 96h; rat: LD50. Experimental toxicities (mol/L) were transformed into the logarithms of the inverse effect/lethal concentrations.	
STRUCTURES	Molecular structures were drawn and minimized by the semi-empirical method AM1 in HYPERCHEM software, and converted into SMILES by Open Babel software [3].	
DESCRIPTORS	Dragon 5.5 , Dragon 6, CADASTER on-line platform [4].	
ALGORITHM	Multiple linear regression (MLR) performed by *Ordinary Least Squares* (OLS) method. Variable selection by Genetic Algorithm (GA).	Multiple linear regression (MLR) performed by *Partial Least Squares* (PLS).
TOOLS of VALIDATION	Internal stability verified by R^2, Q^2_{LOO}, Q^2_{BOOT}, R^2/Q^2_{YS} and RMSE. External predictivity was measured using different Q^2_{EXT} parameters [5-7]. Prediction sets were obtained by splitting 30% (K-ANN , splitting random by response). A blind external prediction set was also used in the fish model.	R^2, Q^2, RMSEE and RMSEP (Root Mean Square Error of Estimation and Prediction).
STRUCTURAL DOMAIN	Leverage (Distance to model in the X space)	DModX (Distance to Model in X space)

FISH TOXICITY

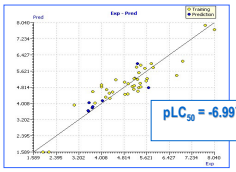
UI: MLR-OLS

- Model based on 39 BTAZs
- External validation on three prediction sets
- Endpoint pLC50 96h
- Descriptors from Dragon 5.5 and other 2D calculated in the CADASTER online platform

MODEL	N_{TR}	N_P	R^2	Q^2_{LOO}	Q^2_{boot}	Q^2_{EXT} *	$RMSE_{tr}$	$RMSE_{ex}$
Split Rand.	28	11	0.83	0.77	0.69	0.66-0.84	0.43	0.61
Split K-ANN	28	11	0.81	0.73	0.67	0.8-0.88	0.49	0.5
FULL	39	-	0.86	0.82	0.74		0.46	

* Range of different parameters for Q^2_{EXT}

Validation parameters on 8 blind external TAZs
Q2ext* = 0.69-0.86, RMSEext = 0.45



pLC_{50} = -6.9979 + 1.5123 *ClC1* + 13.2813 *Mp* + 0.1881 *H-052*

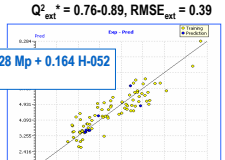
Further enlargement of the dataset with 50 additional azo-aromatic compounds

- Model based on 79 chemicals (29 BTAZs + 50 azo-aromatic comp.)
- 10 BTAZs for common external set
- 8 BTAZs as blind validation set

MODEL	N_{TR}	N_P	R^2	Q^2_{LOO}	Q^2_{BOOT}	Q^2_{EXT} *	$RMSE_{tr}$	$RMSE_{ex}$
Split Rand.	55	24	0.83	0.79	0.75	0.85-0.91	0.52	0.38
Split K-ANN	55	24	0.84	0.81	0.79	0.76-0.88	0.5	0.45
FULL	79	-	0.84	0.81	0.8		0.48	

Validation parameters on
10 common external BTAZs: Q^2_{ext}* = 0.910-0.913 $RMSE_{ex}$ = 0.35
8 blind external BTAZs: Q^2_{ext}* = 0.76-0.89, $RMSE_{ext}$ = 0.39

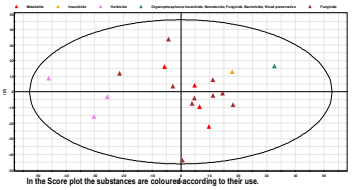
pLC50 = -7.131 + 1.598 ClC1 + 13.928 Mp + 0.164 H-052 -0.005 TPSA(Tot)



- Similar structural information selected by GA
- Nested Model + TPSA(Tot)
- Increase of the model stability and predictivity

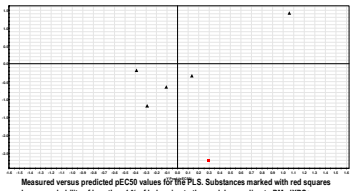
IVL: PLS

- Model based on 19 BTAZs
- Validated with 6 BTAZs
- Endpoint pLC50 96h
- Descriptors from Dragon 6 calculated on the CADASTER online platform



In the Score plot the substances are coloured according to their use.

Distance to the model plane of a prediction is known as DModXPS (Distance to Model in X space for the Prediction Set), while also considering the distance in the model plane leads to the statistic DModXPS+. From these distances and the corresponding distances in the training set, it is possible to calculate a probability that a (new) substance belongs to the model. These probabilities are known as PModXPS and PModXPS+. Outliers in the validation set identified via PModXPS+.



Measured versus predicted pEC50 values for the PLS. Substances marked with red squares have a probability of less than 1 % of belonging to the model according to PModXPS+.

The chemical classified as outlier is removed and a new validation is made. The RMSEP is increasing with more than a factor 2 after the outlier removal.

The model performance is summarized in the table below.

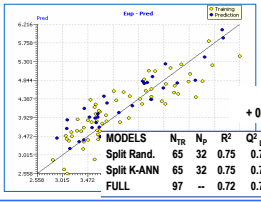
	PModXPS+ 1 %					
Model	R^2	Q^2	RMSEE	RMSEP	outliers	RMSEP [a]
PLS	0.98	0.79	0.18	1.31	1	0.54

[a] RMSEP for the validation set after removal of the outliers indicated by this method.

DAPHNIA TOXICITY

UI: MLR

- Model based on 46 BTAZs + 51 azo-aromatic compounds
- External validation on two prediction sets
- Endpoint pEC50 48h
- 2D Descriptors from CADASTER on-line platform



FULL MODEL equation

pEC_{50} = 3.17 + 0.34 ALogP + 0.11 C(-H'-C'=C') – 0.01 p5-3N

MODELS	N_{TR}	N_P	R^2	Q^2_{LOO}	Q^2_{EXT}	$RMSE_{tr}$	$RMSE_{ex}$
Split Rand.	65	32	0.75	0.71	0.71-0.73	0.40	0.43
Split K-ANN	65	32	0.75	0.71	0.72-0.78	0.41	0.39
FULL	97	--	0.72	0.73	--	0.40	--

IVL: PLS

- Model based on 29 BTAZs
- Validated with 8 BTAZs
- Endpoint pEC50
- Descriptors from Dragon 6

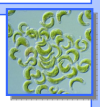
For daphnia 5 of the 8 substances selected for validation were classified as outliers with the same method as for the fish model. The performance of the PLS model is shown in the table below.

	PModXPS+ 1 %					
Model	R^2	Q^2	RMSEE	RMSEP	outliers	RMSEP [a]
PLS	0.97	0.88	0.18	2.33	5	0.37

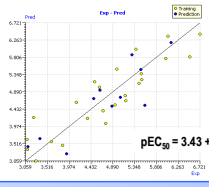
[a] RMSEP for the validation set after removal of the outliers indicated by this method.

ALGAE TOXICITY

UI: MLR



- Model based on 17 BTAZs + 18 azo-aromatic compounds
- External validation on two prediction sets
- Endpoint pEC50 96h
- 2D Dragon descriptors (ver. 5.5)



FULL MODEL equation

pEC_{50} = 3.43 + 0.01 AEigZ + 0.09 T(N..S) + 0.15 SEigv

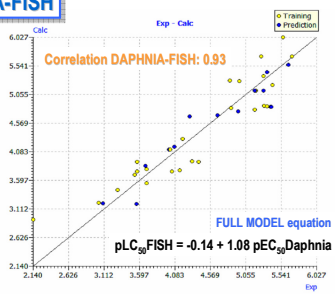
N_{TR}=35 R^2=0.83 Q^2=0.78 RMSE=0.41

SPLIT MODELS	N_{TR}	N_P	R^2	Q^2_{LOO}	Q^2_{EXT}*	$RMSE_{tr}$	$RMSE_{ex}$
Split Rand.	24	11	0.86	0.80	0.72-0.79	0.39	0.47
Split K-ANN	22	13	0.83	0.76	0.74-0.85	0.44	0.42

Interspecies Quantitative Correlation: DAPHNIA-FISH

Interspecies Quantitative Correlations were developed to provide direct estimation of the acute aquatic toxicity of untested BTAZs from *Daphnia* to Fish.

The here proposed linear regression model was obtained by using EC_{50}48h measured in *Daphnia* as independent variable, and LC_{50}96h measured in Fish as response endpoint. The model was externally validated by random splitting of the experimental data set before the development of the model.



Correlation DAPHNIA-FISH: 0.93

FULL MODEL equation

$pLC_{50}FISH$ = -0.14 + 1.08 pEC_{50}Daphnia

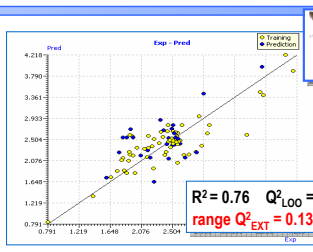
MODELS	N_{TR}	N_P	R^2	Q^2_{LOO}	Q^2_{EXT}*	$RMSE_{tr}$	$RMSE_{ex}$
Split Rand.	27	13	0.85	0.82	0.91-0.93	0.34	0.24
FULL	40	--	0.87	0.85	--	0.31	--

MAMMALIAN TOXICITY

- Data set: 111 BTAZs (large structural and functional heterogeneity)
- External validation on one prediction set (random splitting)
- Endpoint pLD_{50}
- Descriptors from Dragon 5.5

Experimental data mainly collected from ChemID Plus – data quality ?

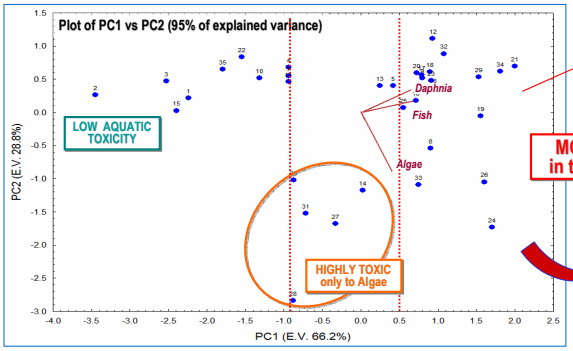
Low external predictivity
→ Internal validation is not sufficient!!



R^2 = 0.76 Q^2_{LOO} = 0.74
range Q^2_{EXT} = 0.13-0.54

BTAZs trend of Aquatic Toxicity by PCA

Available experimental data for EC50 Algae, EC50 Daphnia and LC50 fish were analyzed by PCA in order to characterize the toxicological profile of BTAZs and to identify the most active compounds in the studied scenario.



RANGE OF TOXICITY
Alage: 0.012 – 6.4 mg/L
Daphnia: 0.77 – 10.2 mg/L
Fish: 0.74 – 6.63mg/L

MOST ACTIVE BTAZs in this aquatic scenario

WORK in PROGRESS classification and regression QSARs

TREND OF AQUATIC TOXICITY

CONCLUSIONS

- Different robust and externally predictive QSAR models have been developed to predict the aquatic toxicity of BTAZs in Algae, *Daphnia* and Fish.
- Interspecies Quantitative Correlation – Linear regression model has been developed to predict Fish acute toxicity from *Daphnia* toxicity data. This model has been externally validated and has comparable performances as the QSAR models developed for fish.
- A profile of the aquatic toxicity has been defined for 35 BTAZs by PCA. PC1 separates chemicals globally more toxic, in the analyzed aquatic scenario, from less hazardous BTAZs (Trend of aquatic Toxicity).
- Work is in progress to create QSAR classification and regression models to predict the cumulative aquatic toxicity of BTAZs on the basis of their chemical structures.
- Results obtained from the QSAR modeling of mammalian toxicity highlight the influence of input data quality on QSAR performances, and the importance of external validation to avoid overestimation of predictivity.

REFERENCES

[1] CADASTER FP7 PROJECT - www.cadaster.eu.
[2] Footprint PPDB (Pesticide Properties DataBase), available on-line at: http://sitem.herts.ac.uk/aeru/footprint/en/index.htm.
[3] Open Babel Toolbox, available on-line at: http://openbabel.org/wiki/Main_Page
[4] www.cadaster.eu/database/
[5] Shi L. M.; Fang H.; Tong W.; Wu J.; Perkins R.; Blair R. M.; Branham W. S.; Dial S. L.; Moland C. L.; Sheehan D. M. J Chem Inf Comput Sci 2001, 41, 186-195.
[6] Schüürmann, G.; Ebert R.U.; Chen J.; Wang B.; Kühne R. J Chem Inf Model 2008, 48, 2140-2145.
[7] Consonni, V.; Ballabio, D.; Todeschini, R. J Chem Inf Model 2009,49, 1669-1678.

Financial support, given by European Union through the project CADASTER (FP7-ENV-2007-1-212668), is gratefully acknowledged