CADASTER

CAse studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment

Grant agreement no.: 212668

Collaborative Project

Sub-Priority ENV2007 3.3.1.1:In-silico techniques for hazard-, safety-, and environmental risk-assessment

Task 3.5 New QSARs for relevant end-points (documented models)

Due date of deliverable 3.5: 31 December, 2011
Actual submission date: 22 December 2011

Start date of project: 1 January 2009 Duration: 4 years

Lead Contractor: National Institute of Public Health and the Environment (RIVM), Laboratory for Ecological Risk Assessment
Corresponding authors of document: Paola Gramatica[2], Willie Peijnenburg[1],Magnus Ramberg[3], Tomas Öberg[4], Igor Tetko[5], Nina Jeliazkova[6]

1. RIVM, Laboratory for Ecological Risk Assessment - P.O. Box 1 3720 BA Bilthoven, The Netherlands (marga.deventer@rivm.nl, willie.peijnenburg@rivm.nl)
2. QSAR Research Unit in Environmental Chemistry and Ecotoxicology, University of Insubria, Via J.H. Dunant 3 - 21100 Varese, Italy (paola.gramatica@uninsubria.it)
3. IVL Swedish Environmental Research Institute, Box 210 60, SE- 100 31 Stockholm, Sweden
4. Linnaeus University, School of Natural Sciences, 391 82 Kalmar, Sweden.
5. Helmholtz Zentrum Muenchen - German Research Center for Environmental Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany
6. IdeaConsult Ltd., 4 A.Kanchev str. Sofia 1000, Bulgaria

Deliverable no: 3.5 New QSARs for relevant end-points (documented models)
Nature: Other+Report

Project co-funded by the EU Commission within the Seventh Framework Programme		
Dissemination Level		
PU	Public	X
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

WP 3: Development and validation of QSARs

Work Package Leader: Paola Gramatica (Partner 3: University of Insubria)

Task 3.5 Development of new QSARs

New QSARs for relevant end-points (documented models) (Deliverable 3.5 – month 36)

Overview

The aim of this deliverable is to provide an overview of new local QSAR models, specifically developed according to the OECD principles for QSAR validation, on the chemicals of the 4 CADASTER classes (i.e. Polybrominated diphenyl ethers (PBDEs) (extended to other Brominated Flame retardants (BFRs)), Perfluoroalkylated substances (PFCs) (extended to Poly-fluorinated compounds), Substituted musks/Fragrances, Triazoles/Benzo-triazoles (B)-TAZs)). The modelled endpoints are those found in the open literature and public databases, collected in WP2 and uploaded in the CADASTER database (WP5). As already pointed out in previous reports, even if the focus of the Project is on SIDS endpoints, all the available experimental data, related to the 4 classes, have been modelled (rodent toxicity, various endocrine disruption end points, ecc) in order to prioritize chemicals for their potential hazard and to suggest priority lists of selected compounds to Partners involved in experimental tests (WP2).

The various Partners, involved in WP3, have developed models for the endpoints/classes by applying different modelling approaches. The models were developed taking into account the OECD principles for validation and acceptability of QSARs for regulation purposes, in particular external validation and check of applicability domain.

The models, developed in the Project, have been documented in publications on international journal, peer reviewed (ISI), and in meeting presentations, listed below, and also in the CADASTER database (qspr-thesaurus) and CADASTER website (http://www.cadaster.eu).

Partners involved in WP3:

- Partner 3 (WP Leader): University of Insubria (UI): Paola Gramatica, Ester Papa, Simona Kovarich, Barun Bhhatarai, Mara Luini, Partha Pratim Roy, Stefano Cassani.
- Partner 1: RIVM: Willie Peijnenburg, Marja Wouterse, Erik Steenbergen, Evdert-Jan van de Brandhof
- Partner 4: IVL Swedish Environmental Research Institute: Magnus Rahmberg, Sara Nilsson, Håkan Fridén
- Partner 5: Linnaeus University (LnU): Tomas Öberg, Tao Liu
- Partner 6: Helmholtz Zentrum Muenchen (HMGU): Igor V. Tetko, Stefan Brandmaier, Wolfram Teetz, Iurii Sushko, Faizan Sahigara
- Partner 7: Ideaconsult Ltd.: Nina Jeliazkova, Nikolay Kochev, Ognyan Pukalov

Partner 3: University of Insubria (UI)

QSAR and QSPR models have been developed, according to the OECD principles for the validation for regulatory purposes of (Q)SAR models (http://www.oecd.org/dataoecd/33/37/37849783.pdf) specifically for the four chemical classes, studied in the CADASTER project, i.e. Polybrominated diphenyl ethers (PBDEs) (and Brominated Flame retardants (BFRs)), Perfluoroalkylated substances (PFCs) (and Poly-fluorinated compounds), Substituted Musks/Fragrances, Triazoles/Benzo-triazoles ((B)-TAZs). The models, for all the end points that UI had found in literature with experimentally available data in adequate number and quality for modelling, data which were selected and collected in WP2 and uploaded in WP5, have been developed by Multiple Linear Regression (MLR - OLS method), or classification (*k-nearest neighbours* - k-NN method). The models are based on molecular descriptors, calculated by DRAGON software, PADEL software (open source) and various descriptors available on the CADASTER platform, and then selected by a Genetic Algorithm. Particular attention has been always devoted to the validation aspects: in addition to internal validation (LOO and Bootstrapp), external validation during the model development was always done, when sufficient data were available (>10). Two different splittings on the input data sets have been applied (based on structural similarity by SOM (Kohonen maps) and on response distribution) in order to avoid bias in the selection of the modelling molecular descriptors. Different external validation parameters (Q^2-F1,F2,F3 and Concordance Correlation Coefficient, recently proposed by UI in CADASTER Project[1]) have been always calculated and compared, in order to select only models that are recognized as externally predictive by all the validation criteria. Y-scrambling procedure has been applied to verify the absence of chance correlation in each model. Williams graph has been always verified to highlight outliers for the response and high leverage chemicals. The developed QSAR models have been also applied to hundreds of chemicals, belonging to the four CADASTER classes, without experimental data (the majority in ECHA pre-registration list for REACH): the structural applicability domain (AD) of the proposed models has been verified, in order to label the predicted data as interpolated (more reliable values) or extrapolated (less reliable values) and verify their distribution in an Insubria graph. The structural AD of all the UI models was always demonstrated to be very high (see detailed % below for each model and class). The UI models, with verified high external predictivity and structural applicability for new chemicals of the four CADASTER classes, can be applied by regulators in REACH, with the important information on the reliability of the predicted data (interpolated or extrapolated by the applied model). All the QSAR models were developed and validated by UI using the in house software for model development and validation QSARINS [2] (that will be freely available from the Insubria web: www.qsar.it).This software can calculate all the validation criteria and plot all the graphs for regression and different AD (Williams graph and Insubria graph).

The development of new QSAR models for the studied endpoints started already from the beginning of the project, and not from month 12 as it was originally planned. This was necessary to produce as soon as possible data useful for the prioritisation, according to their physico-chemical properties and

toxicological profiles, of chemicals for experimental testing (WP2) and for the creation and implementation of the QSPR-THESAURUS of QSAR models (WP5).

An interesting result of UI work is also an, externally validated, interspecies relationship model on (B)TAZs aquatic toxicity, that allows to predict toxicity on fish *Oncorhynchus mykiss* starting from experimental data on *Daphnia magna*, thus reducing the number of tests on fish.

The externally predictive QSAR models, developed by UI during the first three years of the project, have been published and documented in papers on international journals peer-reviewed (ISI) (so far 13, plus 3 in preparation) and presented in several international meetings for their wide dissemination. Some on (B)TAZ toxicity are now in preparation. The list of papers and meeting presentations is reported below.

[1.] Nicola Chirico and Paola Gramatica
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., **2011**, *51* (9), pp 2320–2335

[2]Chirico, N.; Papa, E.; Kovarich, S.; Cassani, S.; Gramatica P. ***QSARINS**, software for QSAR model development and validation*, University of Insubria, Varese, Italy, 2011. http://www.qsar.it

Partner 1: RIVM

Activities of RIVM on the development of QSARs were restricted to the initial development of predictive models for perfluorinated compounds (PFCs), based on experimental data collected in the laboratory of RIVM within WP2. The models developed so far require additional improvement and validation. This is foreseen to take place in the final year of the CADASTER project. The current are based on the number of carbon atoms present in the fluorinated alky chain of the chemicals studied. The applicability domain is limited to perfluorinated alkanoic acids and alcohols, further extension will include extension of the applicability domain by generating toxicity data for perfluorinated compounds with additional functionalities.

In addition to these QSARs, interspecies relationships were derived that allow for prediction of EC50-values for either *Daphnia Magna* or *Chydorus sphaericus* on the basis of data or predictions of the other, non-tested, cladoceran species. Similar interspecies relationships were developed to substitute missing data on either lettuce (*Lactuca sativa*) or green algae (*Pseudokirchneriella subcapitata*) on the basis of information of the counter-biotic species.

Partner 4: IVL

Focus from IVL in the development of QSAR is on end-points related to aquatic toxicity. Of the four classes of chemical compounds in CADASTER only the B-TAZ group has enough of data for generating good models for aquatic toxicity.

Descriptors used in the modelling of B-TAZ are calculated from the Dragon software, version 6.

The method for modelling is partial least squares (PLS) regression, which is a latent variable regression method. Since no a priori information about variable importance is available, so called auto-scaling was used, i.e. all variables were scaled with the inverse of their standard deviation in the training set and then

centred by subtracting the mean. One of the advantages of latent variable regression methods are that is offers the possibility of outlier detection. It is very important to note that empirical models are not valid outside the domain in which they are trained, i.e. the applicability domain. Outlier diagnostics can be used to estimate whether or not new substances resembles the data in the training set to an extent high enough to provide reliable predictions. If not, the predictions obtained should not be trusted.

Partner 5: Linnaeus University (LnU)

The modelling approach of Linnaeus University has been based on theoretical descriptors computed from molecular structures. The molecular structures have been optimized into low-energy 3D conformations using the software CORINA. The molecular descriptors were generated using the software DRAGON and include constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, BCUT descriptors, topological charge indices, eigenvalue-based indices, functional group counts, and atom-centered fragments. QSPR and QSAR models developed by LnU are regression models, fitted either by bilinear partial least squares regression (PLSR) or by multiple linear regression (MLR), using the software UNSCRAMBLER, SIMCA and MATLAB.

Non significant descriptor variables were assigned zero weight; these variables were identified using a jackknife method for significance testing of the PLSR model parameters during cross-validation. To evaluate local PLSR models, samples were selected based on the Euclidean distances in descriptor space and uniform weighting was applied. All descriptor variables were preprocessed by auto-scaling to zero mean and unit variance. Cross-validation was used to establish the rank of the calibration model (the number of latent variables), and an external test set was used to estimate the prediction error. An adaptive sample weighting scheme was developed and applied to update a previously developed model. The applicability domain of the PLSR models was assessed by the residual standard deviation (the Euclidean distance to the model) and the leverage (the Mahalanobis distance to the calibration objects within the model space). These two distance measures were then used to decide if an object was within the domain of application or not. Here, the 5% significance level was chosen as the limit for the residual standard deviation and the limit for the leverage was set to three times the average leverage for the calibration objects.

Models developed by Linnaeus University during the CADASTER project have been published or are in the process of being published in the scientific literature. In addition, presentations have been given at various scientific conferences.

Partner 6: HMGU

Many HMGU collaborators were strongly involved in different steps of model development as summarized below. People, who led the corresponding activities, are indicated in parentheses. HMGU has contributed to development of methodology to estimate accuracy of predictions for quantitative and qualitative models

(Iurii Sushko), experimental design to select most informative molecules for testing (Stefan Brandmaier) in collaboration with LnU, collection of experimental data e.g., PhysProp and PPDB that were used in model development by the groups (Stefan Brandmaier). The group also participated in development of boiling and melting point of PFC in collaboration with other groups (Wolfram Teetz) as well as developed toxicity models for fish, *Daphnia* (Fazian Sahigara) and algae for TAZ & BTAZ compounds. It also contributed model for Ames set prediction (Iurii Sushko), which covers molecules from all four classes of chemicals considered in the project. Moreover, recently group has developed global models for boiling and melting points which includes all groups of molecules. It will be compared with local models developed specifically for each group of chemicals analyzed in the project.

Partner 7: Ideaconsult

Linear regression QSAR models for Triazoles and Benzo-triazoles (B-TAZ) have been developed. We used 2D molecular structures encoded in SMILES, and calculated descriptors by DRAGON 5.4 2006 software. Several feature selection procedures have been explored, starting from preliminary variable elimination, removing correlated descriptors and following by genetic algorithm by Mobydigs[1] version 1.0 / 2004 software and expert selection. A set of 20 models with different complexity (5,6,7 and 10 variables) is generated and model selection performed, by analysing the trade-off between the model accuracy and over-fitting. Additional models were created by selecting descriptors by expert knowledge. The expert variable selection is guided by the genetic algorithm, and additionally includes functions of the descriptors variables (e.g. logarithm, square root, power of two, etc.). Model performance is further checked by y-scrambling procedure and leave one out validation, as well as against two external dataset, selected for validation by project partners.

[1] Roberto Todeschini, Viviana Consonni, Andrea Mauri, Manuela Pavan, "Mobydigs: Software For Regression And Classification Models By Genetic Algorithms",Chapter 5 in Chemometrics: Genetic Algorithms and Artificial Neural Networks, 2003, Issue: L, Publisher: Elsevier, Pages: 1-32

Polybrominated diphenyl ethers (PBDEs) (and Brominated Flame retardants (BFRs)),

Brominated flame retardants (BFRs) are a class of hydrophobic chemicals that are incorporated in a variety of consumer products (e.g. electronic devices, building materials, textiles, etc..) to increase their fire resistance. Among the large number of BFRs on the market, CADASTER focused the attention on mainly on polybrominated diphenyl ethers (PBDEs), including all the 209 potential congeners. Additionally, other heterogeneous BFRs were studied under the project, e.g. several PBDE metabolites (OH-PBDEs and CH3O-PBDEs), brominated phenols, tetrabromobisphenol-A (TBBPA) and brominated bisphenol A compounds, hexabromocyclododecane (HBCD) and three alternative compounds to

decaBDE (i.e. decabromodiphenyl ethane – DBDE; ethylene bistetrabromo phthalimide – EBTPI; 1,2-bis(2,4,6-tribromophenoxy) ethane – TBE).

Partner 3 UI activity

Local QSAR/QSPR models have been developed for several physico-chemical properties, environmental fate and toxicity endpoints. The selection of the endpoints for QSAR modelling was based on the few experimental data available. The modelled endpoint are listed below and summarised in Table 1:

- melting point (MP)
- vapour pressure (log VP)
- water solubility (log WS)
- Henry's low constant (log H)
- octanol-air partition coefficient (log K_{OA})
- octanol-water partition coefficient (log K_{OW})
- photolysis rate constant (log K_P)
- photolysis half-life (log HL_P)
- endocrine disrupting (ED) potency (different endpoints, i.e. aryl hydrocarbon receptor binding, agonism and antagonism, EROD induction, estrogen receptor agonism and antgonism, androgen receptor antagonism, progesterone receptor antagonism, T4-TTR competition and E2SULT inhibition)

Table 1. Summary of QSAR/QSPR models developed for BFRs (UI).

	Endpoint	**Method**	**SIDS**	**EPI Suite comparison**	**On-line**	**Reference***
Phys-chem	MP	MLR-OLS	x	x	x	*a, b*
	Log WS	MLR-OLS	x	x		*a*
	Log H	MLR-OLS	x	x	x	*a*
	Log VP	MLR-OLS	x	x	x	*a*
	Log K_{OA}	MLR-OLS	x	x	x	*a, b*
	Log K_{OW}	MLR-OLS	x	x	x	*a, b*
Environmental Fate	Log K_P	MLR-OLS	x			
	Log HL_P	MLR-OLS	x			
ED potency	Log RBA	MLR-OLS				*c*
	Log $1/EC_{50}EROD_{ind}$	MLR-OLS				*c*
	Log $1/EC_{50}DR_{ag}$	MLR-OLS				*c*
	Log $1/EC_{50}ER_{ag}$	MLR-OLS				*c*
	Log $1/IC_{50}PR_{ant}$	MLR-OLS				*c*
	Log $T4_{REP}$	MLR-OLS				*c*
	Log $E2SULT_{REP}$	MLR-OLS				*c*
	DR_{ag}	k-NN				*d*
	DR_{ant}	k-NN				*d*
	ER_{ag}	k-NN				*d*
	ER_{ant}	k-NN				*d*
	AR/PR_{ant}	k-NN				*d*

	$T4\text{-}TTR_{comp}$	k-NN				*d*
	$E2SULT_{inh}$	k-NN				*d*

* *a)* Papa, E.; Kovarich, S.; Gramatica P., *QSAR Comb. Sci.* (2009) 28, 790-796 ; *b)* Papa, E.; Kovarich, S.; Gramatica P., *Mol. Info. (*2011) 30, 232-240; *c)* Papa, E.; Kovarich, S.; Gramatica, P., *Chem. Res. Toxicol.* (2010) 23, 946-954; *d)* Kovarich, S.; Papa, E.; Gramatica, P., *J Haz. Mat.* (2011) 190, 106-112.

All the models are based on 1 or 2 molecular descriptors calculated by the DRAGON software (ver. 5.5). Models are statistically robust, internally and, when possible, externally validated, and with a verified applicability domain. Models have been applied to predict data for all 243 BFRs studied under the project, always verifying the degree of interpolation/extrapolation of predictions (by leverage approach). Model equations, statistical performances, predictions and information on applicability domain (interpolated/extrapolated predictions) are provided in the respective publications.

Some of the QSPRs (MP, VP, WS, H, logKoa, logKow) were compared with the EPI Suite Estimation programs. As expected, prediction accuracy of the local models developed under CADASTER Project was higher than prediction accuracy obtained by applying the general EPI Suit models.

Models developed for MP, VP, logKoa, logKow have been uploaded in the CADASTER database and are freely available. Additionally, predictions (together with the information on applicability domain) of phys-chem properties and ED potency endpoints for 243 BFRs have been uploaded in the CADASTER database among the “calculated properties”.

Experimental and predicted data available for the different endpoints related to ED potency have been used for the prioritization of chemicals (Deliverable 3.4), in order to focus the experimental testing.

The lack of experimental data on eco-toxicity endpoints for brominated flame retardants prevented the development of specific local QSAR models to be used for the environmental risk assessment of this class of chemicals.

Partner 5 LnU activity

A vapour pressure model for polybrominated diphenyl ethers (PBDE), meeting the OECD requirements, was previously developed and published by the LnU group.[1] This model was based on experimental data from the literature and the performance can be summarised as follows: SEC 0.16 (log Pa), R2cal 0.992, SEP 0.13 (log Pa), and Q2ext 0.994.

QSAR models for the bioconcentration of PBDEs are now in development by LnU and a partner in the ECO project. This modelling is primarily based on experimental data generated in the CADASTER WP2.

1. Tomas Öberg
Prediction of vapour pressures for halogenated diphenyl ether congeners from molecular descriptors. *ESPR - Environmental Science and Pollution Research* 9, 405-411 (2002).

Partner 6 HMGU activity

It has developed global models for boiling and melting points, which is compared with the local models developed in the project (under preparation). A model to predict AhR binding activity using Molecular Field Topology Analysis (MFTA) methodology is also under preparation now.

Table 2. Summary of QSAR/QSPR models developed for BFRs (HMGU).

	Endpoint	Method	SIDS	Reference*
Phys-chem	MP	ASNN	x	*Under preparation*
	BP	ASNN	x	*Under preparation*
ED potency	AhR binding	PLS		*Under preparation*

Perfluoroalkylated substances (PFCs) (and Poly-fluorinated compounds)

Per- and polyfluorinated compounds (PFCs) are a class of synthetic substances widely used in different materials as waterproof fabrics, food packaging, non-adhesives, fire-fighting foams, paints, etc.. The amphiphylic nature of some PFCs, characterized by an hydrophobic fluorinated alkyl chain and a polar terminal group (such as carboxylic and sulfonic acids), gives them a great stability, thermal and stress resistance, and excellent surfactant properties.

PFCs studied under the CADASTER Project include 382 chemicals, both linear and aromatic chemicals, with different carbon chain length, fluorination degree (per- and polyfluorinated compounds) and functional groups (carboxylates, sulfonates, sulfonamides, alcohols, etc.).

The majority of these compounds are included in the ECHA pre-registration list.

Partner UI activity

Local QSAR/QSPR models have been developed for several physico-chemical properties and toxicity endpoints. The selection of the endpoints for QSAR modelling was based on the few experimental data available. The modelled endpoint are listed below and summarised in Table 1:

- melting point (MP)
- boiling point (BP)
- vapour pressure (log VP)
- water solubility (log WS)
- critical micelle concentration (log CMC)
- mammalian (rat/mause) acute toxicity (oral/inhalation) (log 1/ LD_{50})
- T4-TTR competing potency (T4-TTR_{comp})

Table 3. Summary of QSAR/QSPR models developed for PFCs (UI).

	Endpoint	Method	SIDS	EPI Suite	On-line	Reference*

				Comparison		
Phys-chem	MP	MLR-OLS	x	x	x	*a*
	BP	MLR-OLS	x	x	x	*a*
	Log VP	MLR-OLS	x	x	x	*b*
	Log WS	MLR-OLS	x	x	x	*b*
	Log CMC	MLR-OLS				*b*
ED potency	LD_{50} (mouse-or)	MLR-OLS				*c*
	LD_{50} (rat-or)	MLR-OLS	x			*c*
	LC_{50} (mouse-inh)	MLR-OLS				*d*
	LC_{50} (rat-inh)	MLR-OLS	x			*d*
	T4-TTR_{comp}	k-NN				*e*

* *a)* Bhhatarai, B. et al. (WP3 partners), *Molecular Informatics* (2011) 30, 189-204; *b)* Bhhatarai, B.; Gramatica, P., *Environ. Sci. Technol.* (2010) 45 (19), 8120-8128; *c)* Bhhatarai, B.; Gramatica, P. , *Molecular Diversity* (2011) 15, 467-476; *d)* Bhhatarai, B.; Gramatica P., *Chem. Res. Toxicol.* (2010) 23, 528-539; *e)* Kovarich, S., Papa, E. Gramatica P., (2011), *SAR QSAR Environ. Res.* (proceedings of CMTPI 2011), in press.

All the models are based on 1 or 2 molecular descriptors calculated by the DRAGON software (ver. 5.5). Models are statistically robust, internally and, when possible for the number of available data, externally validated, and with a verified applicability domain. Models have been applied to predict data for more than 250 PFCs studied under the project (many of them included in the ECHA pre-registration list), always verifying the degree of interpolation/extrapolation of predictions (by leverage approach). Model equations, statistical performances, predictions and information on applicability domain (interpolated/extrapolated predictions) are provided in the respective publications. In particular, 90.9% of the predictions for 376 PFCs were interpolated by the models on rodent oral toxicity, while about 76-77% of 250 PFCs were interpolated by the models on rodent inhalation toxicity.

QSPRs models developed for MP, BP, VP and WS were compared with the EPI Suite Estimation programs. As expected, prediction accuracy of the local models developed under CADASTER Project was higher than prediction accuracy obtained by applying the general EPI Suite models.

As an example:

Response	**RMSE EPISuite**	**RMSE UI model**
Water solubility	1.98	0.96
Vapor pressure	1.13	0.95

Models developed for BP, VP and WS have been uploaded in the CADASTER database and are freely available. Additionally, predictions (together with the information on applicability domain) of phys-chem properties for 382 PFCs have been uploaded in the CADASTER database among the “calculated properties”.

Experimental and predicted data available for mammalian acute toxicity have been used for the prioritization of chemicals (Deliverable 3.4), in order to focus the experimental testing.
The lack of experimental data on eco-toxicity endpoints for PFCs prevented the development of specific local QSAR models to be used for the environmental risk assessment of this class of chemicals.

Partner 1 RIVM activity

Aquatic toxicity prediction for daphnids, lettuce and algae

The following models were developed:

A – Lettuce (*Lactuca sativa)* – endpoint: inhibition of root elongation expressed as EC50 (mM) after 5 days of exposure, nC = number of carbon atoms in the alkyl chain of the compounds tested:

$\text{Log } EC_{50,\text{ lettuce}} = -0.17\ (\pm 0.04) \times nC + 1.2\ (\pm 0.27)$
$n = 5$, $R^2 = 0.853$, $p = 0.0252$

B – Acute toxic effects of PFCs on the photosynthesis of green algae (*Pseudokirchneriella subcapitata*) expressed as EC50 (mM) after 4.5 hours of exposure. nC = number of carbon atoms in the alkyl chain of the compounds tested:

$\text{Log } EC_{50,\text{ algae}} = -0.16\ (\pm 0.01) \times nC + 1.313\ (\pm 0.09)$
$n = 4$, $R^2 = 0.988$, $p = 0.006$

C– Cladoceran (waterflea) *Daphnia magna* – endpoint: immobilisation expressed as EC50 (mM) after 24 and 48 hours of exposure, nC = number of carbon atoms in the alkyl chain of the compounds tested:

$\text{Log } EC50_{24h} = -0.127\ (\pm 0.009) \times nC + 0.646\ (\pm 0.071)$
$n = 5$, $R^2 = 0.986$, $p = 7.090 \times 10^{-4}$
and

$\text{Log } EC50_{48h} = -0.131\ (\pm 0.011) \times nC + 0.615\ (\pm 0.096)$
$n = 6$, $R^2 = 0.971$, $p = 3.265 \times 10^{-4}$

In these equations, the endpoint of assessment was the concentration at which 50 %EC50

D - Benthic cladoceran species *Chydorus sphaericus* – endpoint: immobilisation expressed as EC50 (mM) after 24 and 48 hours of exposure, nC = number of carbon atoms in the allkyl chain of the compounds tested:

$\text{Log } EC50_{24h} = -0.209\ (\pm 0.024) \times nC + 0.970\ (\pm 0.202)$
$n = 6$, $R^2 = 0.950$, $p = 9.359 \times 10^{-4}$

and

$\text{Log } EC50_{48h} = -0.201\ (\pm 0.039) \times nC + 0.689\ (\pm 0.327)$

$n = 6$, $R^2 = 0.871$, $p = 6.48\times10^{-3}$

In addition to these QSAR, interspecies relationships were derived that allow for prediction of EC50-values for either *Daphnia Magna* or *Chydorus sphaericus* on the basis of data or predictions of the other, non-tested,,cladoceran species:

For 24-h toxicity:

$\text{Log EC50}_{C.\ sphaericus} = 1.6\ (\pm 0.3) \times \log \text{EC50}_{D.\ magna} - 0.1\ (\pm 0.1)$
$n = 5$, $R^2 = 0.888$, $p = 0.016$

For 48-h toxicity:

$\text{Log EC50}_{C.\ sphaericus} = 1.5\ (\pm 0.3) \times \log \text{EC50}_{D.\ magna} - 0.3\ (\pm 0.17)$
$n = 6$, $R^2 = 0.846$, $p = 0.009$

It can be seen that the relationships between the log-transformed EC50 values of the two cladocerans are significant, so the toxicity of a certain PFC for one cladocerans species can be used to predict the toxicity for the other using the equations. *D. magna* is a pelagic species that inhabits the upper water column, whereas *C. sphaericus* is a benthic species that lives on the sediments. Therefore their EC50s represent aqueous and sediment toxicity of a chemical via exposure to the water phase, respectively. With these interspecies relationships, one could calculate aqueous or sediment toxicity of a similar PFC with known sediment or aqueous toxicity data. Furthermore, the Chydotox toxicity test needs less chemicals and materials, so it may be a promising test method for collection of toxicity data that are needed for environmental risk assessment.

Similar interspecies relationships were developed to substitute missing data on either lettuce or algae on the basis of information of the counter-biotic species:

$\text{Log EC}_{50,\ lettuce} = 1.196\ (\pm 0.437) \times \log \text{EC}_{50,\ algae} - 0.245\ (\pm 0.161)$
$n = 4$, $R^2 = 0.789$, $p = 0.11$

1. G. Ding, M. Wouterse, R. Baerselman, W.J.G.M. Peijnenburg. Toxicity of poly- and perfluorinated compounds to lettuce (*Lactuca sativa*) and green algae (*Pseudokirchneriella subcapitata*). Arch. Environ. Sci. Technol., accepted for publication, 2011.
2. G. Ding, E.-J. van den Brandhof, R. Baerselman, W.J.G.M. Peijnenburg. Acute toxicity of poly- and perfluorinated compounds to two cladocerans, *Daphnia magna* and *Chydorus sphaericus*. Environ. Toxicol. Chem., accepted for publication, 2011.

Partner 5 LnU activity

LnU has extended the applicability domain of a general a QSPR model for vapour pressure[1] to include perfluorinated alkylated substances. Only a few reliable measurements of vapour pressure for perfluorinated carboxylic acids and fluorotelomer alcohols were identified in the literature (n=11). The re-calibration was accomplished by including three of these in the updated model with a sample weight of 14. The re-fitted PLSR model was subsequently evaluated with the remaining 8 compounds as an external test set (Q2ext 0.994 and RMSEP 0.199 log Pa). A comparison with predictions by SPARC and EPI Suite models showed that these models have a substantial systematic bias and overpredicts the vapour pressure. Such a systematic bias was virtually nonexistent for the recalibrated QSPR model and the precision was also improved. The model was subsequently applied to a large set of perfluorinated substances and could predict vapour pressure for more than 200 compounds were reliable experimental data are missing. The recalibrated model has been published and is available on the CADASTER web. In collaboration with other WP3 partners, LnU has also developed QSPR models for other perfluorinated compounds to estimate melting and boiling points. The PLSR models developed by LnU are reported in a joint paper with the other collaborators.

1. Tomas Öberg, Tao Liu
Global and local PLS regression models to predict vapor pressure
QSAR & Combinatorial Science 27, 273-279 (2008).

Partner 6 HMGU activity

Participated to development of collaborative models for boiling and melting point of PFC compounds. We have also developed global models for these properties, which will be compared with the local models in a publication, which is under preparation now.

Table 4. Summary of QSAR/QSPR models developed for PFCs (HMGU).

Endpoint	Method	SIDS	Reference*
MP	ASNN	x	*Under preparation*
BP	ASNN	x	*Under preparation*

a) Bhhatarai, B. et al. (WP3 partners), *Molecular Informatics* (2011) 30, 189-204;

Fragrances

Substituted musks/fragrances are a heterogeneous group of chemicals of varying composition. Examples include substituted benzophenones, polycyclic musks, salicylates, cinnamates and other esters with fragrances behaviour, and terpene derivatives. In view of their typical use pattern, the chemicals have a common emission pattern in the environment.

Fragrances studied and modelled so far under the CADASTER Project include 146 chemicals, selected in the literature and uploaded in the web by Partner 3 (UI) or provided by RIFM, belonging to different chemical classes.

Partner 3 UI activity

Local QSAR/QSPR models have been developed for the following physico-chemical properties and toxicity endpoints:

- vapour pressure (log VP)
- water solubility (log WS)
- octanol-water partition coefficient (log K_{OW})
- acute toxicity in mouse (log $1/LD_{50}$ oral)
- cyto-toxicity in rat (Log EC_{50} NADH-Ox, Log EC_{50} Dψm)

The selection of the endpoints for QSAR modelling was based on the few experimental data available. The modelled endpoint are summarised in Table 2:

Table 5. Summary of QSAR/QSPR models developed for Fragrances (UI).

	Endpoint	**Method**	**SIDS**	**EPI Suite comparison**	**Reference***
Phys-chem	Log VP	MLR-OLS	x	x	*a*
	Log WS	MLR-OLS	x	x	*a*
	Log K_{OW}	MLR-OLS	x	x	*a*
Toxicity	Log $1/LD_{50}$	MLR-OLS	x		*b*
	Log EC_{50} NADH-Ox	MLR-OLS			*b*
	Log EC_{50} Dψm	MLR-OLS			*b*

* *a)* Papa, E.; Luini, M.; Gramatica, POSTER presented at SETAC-Europe 2009; b) Papa, E.; Luini, M.; Gramatica, P. *SAR QSAR Environ. Res.* (2009) 20, 767–779.

All the models are based on 2 or 3 molecular descriptors calculated by the DRAGON software (ver. 5.5). Models are statistically robust, internally and externally validated, and with a verified applicability domain. Models have been applied to predict data for 79 fragrances, always verifying the degree of interpolation/extrapolation of predictions (by leverage approach). Model equations, statistical performances, predictions and information on applicability domain (interpolated/extrapolated predictions) are provided in the respective publications.

Predictions obtained by the local QSPR models developed for VP, WS, and logKow were compared with predictions calculated by EPI Suite models. VP and logKow local models are characterized by higher prediction accuracy than the general EPI Suit models.

Models developed for phys-chem properties are now under revision.

Experimental and predicted data available for mammalian toxicity and cyto-toxicity have been used for the prioritization of chemicals (Deliverable 3.4), in order to focus the experimental testing.
QSAR model for the prediction of ready biodegradation of fragrances are now under development and will be validated with experimental data tested under CADASTER Project (WP2).

Partner 6 HMGU activity

HMGU has developed global models for boiling and melting points, the results of which are compared to predictions made with local models (manuscript under preparation).

Table 6. Summary of QSAR/QSPR models developed for Fragrances (HMGU).

	Endpoint	**Method**	**SIDS**	**Reference***
Phys-chem	MP	ASNN	x	*Under preparation*
	BP	ASNN	x	*Under preparation*

Triazoles and Benzotriazoles (B)-TAZs))

Triazoles and benzotriazoles (B-TAZs) are a class of synthetic molecules characterized by the presence of a simple or condensed aromatic heterocyclic ring (2C + 3N atoms). (B)TAZs find a wide application in many fields; they are used as components of many pesticides, pharmaceuticals (e.g. painkillers, antimycotic and antidepressants medicines), UV stabilizer for plastics, but also they are abundantly used as components of liquid de-icing agents for aircraft and airport runways.
(B)TAZs studied under the CADASTER Project include 386 compounds, included also in the ECHA pre-registration list, structurally highly heterogeneous, and characterized by different using pattern and mechanism of actions.

Partner 3 UI activity

Local QSAR/QSPR models have been developed for several physico-chemical properties and eco-toxicity endpoints. The selection of the endpoints for QSAR modelling was based on the available experimental data, which were mainly collected from freely available databases (e.g. SRC PhysProp database, PPDB Footprint Database). The modelled endpoint are listed below and summarised in Table 1:

- melting point (MP)
- vapour pressure (log VP)
- water solubility (log WS)
- octanol-water partition coefficient (log K_{OW})

- acute toxicity in fish (LC50 96h, *Oncorhynchus mykiss*)
- acute toxicity in aquatic invertebrates (EC50 48h, *Daphnia magna*)
- acute toxicity in algae (EC50 72h, *Pseudokirchneriella subcapitata*)
- terrestrial toxicity (LD50 earthworms, honeybees, birds)

Table 7. Summary of QSAR/QSPR models developed for B-TAZs (UI).

	Endpoint	**Method**	**SIDS**	**EPI Suite**	**On-line**	**Reference***
Phys-chem	MP	MLR	x	x		*a*
	Log VP	MLR	x	x		*a*
	Log WS	MLR	x	x	x	*a*
	Log Kow	MLR	x	x	x	*a*
Ecotoxicity	EC_{50} (algae)	MLR	x	x		*Paper in preparation*
	EC_{50} (Daphnia)	MLR	x	x		*Paper in preparation*
	LC_{50} (fish)	MLR	x	x		*Paper in preparation*
	LD_{50} (worms)	K-NN	x			*Under development*
	LD_{50} (bees)	K-NN	x			*Under development*
	LD_{50} (birds)	K-NN	x			*Under development*

* *a)* Bhhatarai B., P. Gramatica, *Water Res.* (2011) 45, 1463-1471.

Models developed for predicting physico-chemical properties are based on 3 or 4 Dragon molecular descriptors (1D, 2D, 3D). Models developed for eco-toxicity endpoints are base on mono- and bi-dimensional descriptors calculated using three different programs: the commercial software Dragon (version 5.5), the CADASTER on-line platform and the freely available software Padel. In this case separate models were developed for each group of descriptors, and they were identified as “Dragon Model”, “CADASTER Model” and “Padel Model” respectively.

Models are statistically robust, internally and externally validated, and with a verified applicability domain. Models have been applied to predict data for more all the 386 B-TAZs studied under the project (many of them included in the ECHA pre-registration list), always verifying the degree of interpolation/extrapolation of predictions (by leverage approach).

Equations, statistical performances, predictions and information on applicability domain (interpolated/extrapolated predictions) for the models developed for the four phys-chem properties are provided in the publication (Bhhatarai and Gramatica, 2011). All the models, when compared with the respective models available in EPI Suite, showed a better accuracy in prediction (a comparable prediction accuracy was found only for the logKow model).

Models developed for VP and logKow have been uploaded in the CADASTER database and are freely available. Additionally, predictions (together with the information on applicability domain) of the modelled phys-chem properties for 386 B-TAZs have been uploaded in the CADASTER database among the "calculated properties".

An interspecies relationship model has been developed on 40 BTAZs: this model allows to predict toxicity on fish *Oncorhynchus mykiss* (log1/LC50) starting from experimental data on *Daphnia magna* (log1/EC50). The relationship has an high possibility of generalization , being based on a wide information (40 compounds) and more importantly being *a priori* verified for its external predictivity: the model, developed on 27 BTAZs, is able to predict 13 external BTAZs with Q2 ext= 0.92.

The UI-models (OLS) for aquatic toxicity have been presented in several international meetings in 2011 and relative papers are now in preparation, while k-NN classification models for terrestrial toxicity are now in progress. Equations, statistical performances, predictions and information on structural applicability domain to chemicals without data (interpolated/extrapolated predictions) of these models will be provided in the respective publications. Once published, all the developed models (if possible) and predictions will be uploaded in the CADASTER database.

Partner 4 IVL activity

QSAR models have been developed for aquatic toxicity for three different species; *Onchorhynchus mykiss, Daphnia magna, Pseudokirchneriella subcapitata,* statistics are shown in table 8 below.

Descriptors used in the modelling were calculated by the Dragon software, version 6.0, including 0D, 1D, 2D and 3D descriptors.

When PLS regression is used as modelling method, as done by IVL, two measures can be considered to determine if a new substance is in the model domain, i.e. applicability domain. The first is the distance to the model plane (also called residual magnitude) and the second is the distance between the model centre and the projection in the model plane. In the SIMCA software, used for PLS, the 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+, respectively, in the software. These probabilities can be used for classifying outliers in the PLS models. Since the models will be used by the CADSTER website, algorithms for calculating these probabilities outside the SIMCA software have been developed. They will be incorporated in the CADASTER database in close collaboration with WP5.

Table 8. PLS models developed for B-TAZs

Species	**Method**	**N_{TR}**	**N_P**	**R^2**	**Q^2**	**RMSEE**	**RMSEP**	**outliers**	**RMSEP [a]**	**Reference**
Onchorhynchus	PLS	24	19	0.98	0.79	0.18	1.31	1	0.54	Paper in

mykiss										preparation
Onchorhynchus mykiss[*]	PLS	78	19	0.86	0.79	0.45	0,39	2	0.39	Paper in preparation
Daphnia magna	PLS	33	8	0.97	0.88	0.18	2.33	5	0.37	Paper in preparation
Pseudokirchneriella subcapitata	PLS	15	0	0.99	0.84	0.16	-	-	-	Paper in preparation

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

[*]Original dataset expanded with 49 triazins.

For the algae, *Pseudokirchneriella subcapitata*, all available data was used in the training set. We tried to split the training set but it resulted in poor model performance so it was decided to include all data. This resulted in that the RMSEP value could not be calculated but only the Root Mean Square Error of the Estimation, RMSEE, the fit for observations in the training set were reported. New data from WP2 will be used for external validation when available.

Prior to the PLS modelling a PCA (Principal Component Analysis) were performed. The PCA were based on the 386 compounds specified above. However, aquatic toxicity data were not available for all of these compounds thus the actual PLS models developed were based on considerable fewer substances.The substances to be included in the models were chosen from the PCA to span a maximal descriptor space. Hence some of the substances with toxic endpoints were removed. The actual numbers of chemicals for each model are presented in table 8. To expand the chemical domain, other substances not belonging to the B-TAZ group were introduced. The introduced substances have to be similar in the x-space (descriptor space) otherwise they will be classified as outliers. The result did not in any significant way increase the predictive power, for the endpoints *Daphnia magna* and the algae of the models. For the fish model the RMSEP value decreased and the model performance are reported in table 8 as *Onchorhynchus mykiss**. The expansion will continue if/when other data sets with the same end-points are presented.

Partner 5 LnU activity

A QSAR for acute fish toxicity (LD_{50}) was developed using literature data collected in WP2. This data set is heterogeneous, representing several modes of action, and an initial attempt to extend it with narcosis acting compounds was therefore not successful. Instead a general PLSR model was fitted and validated using the same data as UI (a calibration set of 76 triazoles and triazines and a validation set of 18 triazoles). The model is based on 678 theoretical descriptors (Dragon 6) projected down to four latent variables selected by cross-validation. These latent variables can be considered as new meta-descriptors. The performance of this model can be summarised as follows: RMSEC 0.285 (log units), R2cal 0.935, RMSEP 0.478 (log units), and Q2ext 0.814. One organotin compound and two organophosphorous compounds were slightly out of the applicability domain.

Partner 6 HMGU activity

HMGU has developed models for toxicity of fish, daphnia and algae. For each of these properties we calculated 168 models using different combinations of descriptor sets and modelling approaches with the comprehensive modelling framework developed at On-line Chemical modelling Environment. The used methods included:

k Nearest Neighbors (kNN) predicts a property for a compound using the consensus voting of k compounds from the training set that are nearest to it according to some distance metric**. W**e used Euclidean distance calculated using normalized descriptors (mean 0 and standard deviation 1). The number of nearest neighbors that provided the highest accuracy of classification was calculated following a systematic search in range (0, 100).

ASsociative Neural Network (ASNN) uses the correlation between ensemble responses as a measure of distance amid the analyzed cases for the nearest neighbor technique.[1,2] Thus ASNN performs kNN in the space of ensemble predictions. This provides an improved prediction by the bias correction of the neural network ensemble. The configurable options are: the number of neurons in the hidden layer, the number of iterations, the size of the model ensemble and the method of neural network training. The default values provided at OCHEM web site were used.

Fast Stagewise Multivariate Linear Regression (FSMLR) is a procedure for stage-wise building of linear regression models by means of greedy descriptor selection.[3]

Partial Least Squares (PLS). The number of latent variables was optimized automatically using 5-fold cross-validation on the training set.

Multiple Linear Regression Analysis (MLRA) uses step-wise variable selection. The method eliminates on each step one variable that has regression coefficient non-significantly different from zero (according to the *t*-test). Thus MLRA has only one parameter, ALPHA, which corresponds to the *p*-value of variables to be kept for the regression. ALPHA=0.05 was used.

Support Vector Machine (SVM) uses the LibSVM program. The SVM method has two important configurable options: the SVM type (ε-SVR and μ-SVR) and the kernel type (linear, polynomial, radial basis function and sigmoid). Classic ε-SVR and radial basis function kernel were used. The other options were optimized using default grid search.

The analysed descriptor sets and the related references are reported in Appendix IV.

Moreover, we explored different filtering options. The models, which provided the highest prediction accuracy were identified for each dataset. These models were used to build the consensus models for each property.

Table 9. Summary of QSAR/QSPR models developed for B-TAZs (HMGU).

	Endpoint	Method	SIDS	Reference*
Phys-chem	MP	ASNN	x	*Under preparation*
	EC_{50} (algae)	ASNN	x	*Paper in preparation*
Ecotoxicity	EC_{50} (Daphnia)	ASNN	x	*Paper in preparation*
	LC_{50} (fish)	ASNN	x	*Paper in preparation*

Partner 7 IdeaConsult activity.

MLR models developed for LC_{50}-96h (fish, *Oncorhynchus mykiss*) of (B)TAZs.

NTr.= 76 (28 (B)TAZs + 48 Other Azo-Aromatic compounds)

Test set EV1=10 (B)TAZs
Validation set EV2= 8 (B)TAZs

Molecular structures encoded in SMILES notation were used. Calculation of molecular descriptors was performed using DRAGON 5.4 software. The final number of calculated descriptors was 929. Two filtering criteria were then applied as preliminary variable selection procedure. The first filter is pairwise correlation which removes one of each pair of highly correlated (R > 0.9) descriptors. The second criteria removes constant and near constant descriptors. The final set of molecular descriptors (n=721) was used as input variables for MobyDigs program which performs more elaborate variable selection procedure by applying genetic algorithm.

The genetic algorithm was carried out by using the following *tabu* list criteria:

- R2(x,y) > 0.01
- Correlation between x/x < 0,95
- Standardized entropy > 0.05

If any descriptor violates one of the above conditions it was send to a t*abu* list, i.e. it was not used in the model developing process. Q2 (LOO cross-validation correlation coefficient) was used as fitness function and a maximum six variables in developed models were allowed.

The best 10 models (with 5 and 6 variables) were selected and their performance further checked by y-scrambling procedure and bootstrap validation (both were set up to 1000 iterations) as well as their performance against validation test dataset. Models with 7 and 10 variables are generated by a similar procedure. Additionally, several models (Model1 - Model3) were created by choosing the descriptors by expert selection. This selection was based on the variables obtained by means of genetic algorithm, including also the modified versions of these variables (i.e. descriptors or functions of these descriptors).

Details of the selected descriptors and the statistical quality of the models are reported in Appendix IV.

Papers on local models developed in CADASTER Project for CADASTER classes or general models specifically applied to CADASTER chemicals.

1) Barun Bhatarai and Paola Gramatica
Per- and Poly-fluoro Toxicity (LC50 inhalation) Study in Rat and Mouse using QSAR Modeling.
*Chemical Research in.Toxicol*ogy, **2010,** 23(3), 528-539.

2) Barun Bhhatarai and Paola Gramatica
Oral LD_{50} Toxicity Modeling and Prediction of Per- and Polyfluorinated Chemicals on Rat and Mouse,
Molecular Diversity, **2011**, 15 (2), 467-476.

3) Barun Bhhatarai and Paola Gramatica
Predicting physico-chemical properties of emerging pollutants: QSPR modeling of Benzo(triazoles),
Water Research, **2011**, 45 (3) 1463-1471.

4) Barun Bhhatarai and Paola Gramatica
Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals,
Environmental Science & Technology, **2011**, 45(19), 8120-8128.

5) Bhhatarai, Barun; Teetz, Wolfram; Liu, Tao; Oberg, Tomas; Jeliazkova, Nina; Kochev, Nikolay; Pukalov, Ognyan; Tetko, Igor; Kovarich, Simona; Papa, Ester; Gramatica, Paola,
CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals
Molecular informatics (proceedings EuroQSAR2010), **2011**, 30 (2-3),. 189-204.

6) Simona Kovarich, Ester Papa and Paola Gramatica
QSAR classification models for the prediction of endocrine disrupting activity of brominated flame retardants,
J.Hazardous Materials, **2011**, 190 (1-3), 106-112

7) Simona Kovarich, Ester Papa, Jiazhong Li, Paola Gramatica
QSAR classification models for the screening of the Endocrine Disrupting activity of perfluorinated compounds,
SAR QSAR Environ Res., proceedings CMTPI 11, **in press.**

8) Tomas Öberg and Tao Liu
Extension of a prediction model to estimate vapor pressures of perfluorinated compounds (PFCs).
Chemometrics and Intelligent Laboratory Systems **2011,**107, 59-64.

9) Ester Papa, Simona Kovarich and Paola Gramatica
Development, Validation and Inspection of the Applicability Domain of QSPR Models for physico-chemical properties of Polybrominated DiphenylEthers
*QSAR & Combinatorial. Sci*ence, **2009**, 28 (8), 790-796.

10) Ester Papa, Simona Kovarich and Paola Gramatica
QSAR modeling and prediction of the endocrine disrupting potencies of brominated flame retardants,
*Chemical Research in Toxicol*ogy **2010**, 23 (5), 946-954.

11) Ester Papa, Simona Kovarich and Paola Gramatica
On the use of local and global QSARs for the prediction of Physico-Chemical Properties of Polybrominated Diphenyl Ethers
Molecular informatics (proceedings EuroQSAR2010), **2011**, 30 (2-3), 232-240.

12) Ester Papa, Mara Luini and Paola Gramatica
QSAR modelling of oral acute toxicity and cytotoxic activity of fragrance materials in rodents
SAR & QSAR in Environmental Research, **2009**, 20 (7–8), 767–779.

13) Partha Pratim Roy, Simona Kovarich, Paola Gramatica
QSAR model reproducibility and applicability: a case study of rate constants of hydroxy radical reaction models applied to Polybrominated Diphenyl Ethers and (Benzo)Triazoles
J. of Computational Chemistry, **2011**, 32, 2386-2396.

GENERAL PAPERS FOR QSAR MODELING IN THE CADASTER PROJECT.

14) Nicola Chirico and Paola Gramatica
Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient
J. Chemical Information and Modeling, **2011**, 51 (9), 2320–2335

15) Tao Liu and Tomas Öberg
Modelling of partition constants: Linear solvation energy relationships or PLS regression?
J. of Chemometrics **2009**, 23, 254-262.

16) Ullrika Sahlin, Monika Filipsson and Tomas Öberg
A risk assessment perspective of current practice in characterizing uncertainties in QSAR regression predictions.
Molecular Informatics **2011,** 30, 551-564. .

Papers on global models which include CADASTER chemicals and allows estimation of the applicability domain of models and their accuracy of predictions

17) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Muller, K. R.; Xi, L.; Liu, H.; Yao, X.; Oberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *J. Chemical Information and Modeling,* **2010,** 50(12), 2094-2111.

18) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A.K.; Kovalishyn, V.V.; Prokopenko, V.V.; Tetko, I.V. Applicability domain for in silico models to achieve accuracy of experimental measurements. *J. Chemometrics.* **2010**, 24(3-4), 202-208.

Presentations of CADASTER models in Meetings.

PRESENTATION OF THE PROJECT :
Willie J.G.M. Peijnenburg, Mojca Durjava, Paola Gramatica, Erik Furusjö, Tomas Öberg, Nina Jeliazkova, Mark A.J. Huijbregts, Mike Comber, Igor V. Tetko, Case studies on the Development and Application of in silico Techniques for Environmental hazard and Risk assessment (CADASTER)

- 13th Int. Workshop on QSARs in the Environmental Sciences (Syracuse, N.Y., USA, 8-12 June 2008). Tetko and Gramatica
- 1st International Workshop “Fluorinated Surfactants: New Developments”, June 26-28, 2008, Idstein, Germany. Tetko
- Workshop Mathematics in Biosciences, July 21-23, 2008, Munich, Germany
- 17th European Symposium on QSARs & Omics Technologies and Systems Biology Uppsala, Sweden. Sep 21-26 2008. Gramatica

- International Symposium on Green Chemistry for Environment and Health, Neuherberg, Germany, October 13-16; 2008. Tetko
- Exemplification of the integration of tools within REACH: the CADASTER project, SETAC Europe annual meeting, Milan, Italy May 15-19, 2011. Platform Peijnenburg

FLAME RETARDANTS

QSPR prediction of physico-chemical properties and degradation of PBDEs

Ester Papa, Simona Kovarich, and Paola Gramatica,

- 18th Annual Meeting SETAC-Europe (Warsaw, Poland, 25-29 May 2008) platform Papa
- 13th Int. Workshop on QSARs in the Environmental Sciences (Syracuse, N.Y., USA, 8-12 June 2008) Platform Gramatica
- XI° Congr. Naz. di Chimica dell'Ambiente e dei Ben i Culturali (SCI), Muggia (Trieste, I), 16-20/6/2008. Platform Papa

QSAR prediction of endocrine disruption potencies of brominated flame retardants

Ester Papa, Simona Kovarich, and Paola Gramatica,

- 18th Annual Meeting SETAC-Europe (Warsaw, Poland, 25-29 May 2008)
- 13th Int. Workshop on QSARs in the Environmental Sciences (Syracuse, N.Y., USA, 8-12 June 2008)
- XI° Congr. Naz. di Chimica dell'Ambiente e dei Ben i Culturali (SCI), Muggia (Trieste), 16-20 Giugno 2008.
- SETAC-Europe 2009 - 19th Annual Meeting SETAC-Europe, Göteborg, Sweden, 31 May – 4 June 2009
- CMTPI-2009 - Fifth International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources, 4-8 July Istanbul, Turkey

QSPR prediction of physico-chemical properties and endocrine disruption activity of brominated flame retardants

Ester Papa, Simona Kovarich and Paola Gramatica,

- 17th European Symposium on QSARs & Omics Technologies and Systems Biology Uppsala, Sweden. Sep 21-26 2008
- SETAC North America 30th Annual Meeting, New Orleans, USA, 19 - 23 November 2009, platform Gramatica

Classification QSAR Models for the prediction of endocrine disruption potencies of brominated flame retardants

Simona Kovarich, Paola Gramatica, Ester Papa

- SETAC-Europe 2010 - 20th Annual Meeting SETAC-Europe, Seville, Spain, 23-27 May 2010.
- 14th International Worskshop on QSAR in Environmental and Health Sciences, Montreal, Canada, 24-28 May 2010.

CADASTER Models for Brominated Flame Retardants.

Ester Papa, Simona Kovarich and Paola Gramatica,
18th European Symposium on QSARs (EuroQSAR2010) Rhodes (Greece)19-24 Sept. 2010

PFCs

Rodent toxicity QSAR studies of perfluoro- compounds

Bhhatarai B.; Gramatica P.

- ICCE 2009 - 12th EuCheMS International Conference on Chemistry & the Environment, Stockholm, Sweden, 14-17 June 2009.
- SETAC North America 30th Annual Meeting, New Orleans, USA, 19 - 23 November 2009.
- 14th International Worskshop on QSAR in Environmental and Health Sciences, Montreal, Canada,

24-28 May 2010.
- 18th European Symposium on QSARs (EuroQSAR2010) Rhodes (Greece) 19-24 Sept 2010, invited Plenary lecture of Prof. Paola Gramatica

QSPR Studies for predicting physico-chemical properties of perfluorinated compounds
Barun Bhhatarai, Paola Gramatica, CC2009: Conferentia Chemometrica 2009, 27-30 September, Siofok (Hungary), Platform presentation Dr. Bhhatarai.

QSAR prediction of the Endocrine Activity of perfluorinated compounds,
Simona Kovarich, Ester Papa, Paola Gramatica,
- SETAC-Europe 2010 - 20th Annual Meeting SETAC-Europe, Seville, Spain, 23-27 May 2010.
- EuChems, ICCE 2011, Zurich (CH), 11/9/2011.

QSPR Models for Predictions and Data Quality Assurances: Melting Point and Boiling Point of Perfluorinated Chemicals
Bhhatarai B., Teetz W., Öberg T., Liu T., Jeliazkova N., Kochev N., Pukalov O., Tetko I., Gramatica P.
- 2nd international workshop on new developments of fluorinated surfactants. Idstein, Germany, Jun 17-19, 2010.
- 18th European Symposium on QSARs (EuroQSAR2010) Rhodes (Greece)19-24 Sept 2010.

Updating existing QSAR models - selection and weighting of new data
Tomas Öberg and Tao Liu
- 5th German Conference on Chemoinformatics in Goslar, November 8-10, 2009.

FRAGRANCES

Chemometrical approaches for the characterization of the environmental behaviour of fragrances,
Papa E.; Luini M.; Gramatica P.
SETAC-Europe 2009 - 19th Annual Meeting SETAC-Europe, Göteborg, Sweden, 31 May – 4 June 2009

MORE CADASTER CLASSES

QSAR modelling of toxicity endpoints of emerging pollutants: Fragrances and Perfluorinated compounds
Barun Bhhatarai, Paola Gramatica, Mara Luini, Ester Papa, CMTPI-2009 - Fifth International Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources, 4-8 July Istanbul, Turkey. Major talk by Gramatica

QSAR prediction of physico-chemical properties and biological activities of emerging pollutants: brominated flame retardants and perfluorinated-chemicals,
Paola Gramatica, Barun Bhhatarai, Simona Kovarich and Ester Papa,
Sixth Indo-US Workshop on Mathematical Chemistry, Kolkata (India), 8-10 January 2010, platform Gramatica.

QSAR and QSPR models for emerging pollutants: WP3 activities within the FP7 European Project CADASTER
Simona Kovarich, Barun Bhhatarai, Ester Papa, Magnus Rahmberg, Sara Nilsson, Tao Liu, Tomas Öberg, Nina Jeliazkova, Nikolay Kochev, Ognyan Pukalov**,** Wolfram Teetz, Stefan Brandmaier, Igor V. Tetko, Paola Gramatica
- SETAC-Europe 2011 - 21th Annual Meeting SETAC-Europe, Milan, Italy, 15-19 May 2011
- 6th Int.Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2011), Maribor, Slovenia 3th-7th September 2011.

Physico-chemical property prediction of emerging pollutants: PFC and (B)TAZ for environmental distribution
Barun Bhhatarai, Paola Gramatica

- SETAC-Europe 2011 - 21th Annual Meeting SETAC-Europe, Milan, Italy, 15-19 May 2011.
- 6th Int. Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2011), Maribor, Slovenia 3th-7th September 2011. Platform Bhhatarai
- 16th Int. Symposium on Environ. Pollution and its impact on Life in the Mediterranean region (MESAEP), Ioannina (Greece), 24-27 Sept. 2011.

Exploring the QSARs for OH Tropospheric Degradation of VOCs using freely available online descriptors: application to PBDEs and (B)TAZs
Partha Pratim Roy, Simina Kovarich, Ester Papa, Paola Gramatica,

- SETAC-Europe 2011 - 21th Annual Meeting SETAC-Europe, Milan, Italy, 15-19 May 2011
- 6th Int. Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2011), Maribor, Slovenia 3th-7th September 2011. Poster
- Conferentia Chemometrica 2011, Sümeg, Hungary, September 18-21, 2011

Predictive QSAR modelling for Screening and Prioritization of Environmental Organic Pollutants
P.Gramatica
6th Int. Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2011), Maribor, Slovenia 3th-7th September 2011. Plenary invited Lecture Gramatica

TRIAZOLES AND BENZOTRIAZOLES
QSAR Prediction of Aquatic Toxicity of Triazoles and Benzo-Triazoles
Cassani, S.; Kovarich, S.; D'Onofrio, E.; Papa, E.; Roy, P.P.; Gramatica P.

- SETAC-Europe 2011 - 21th Annual Meeting SETAC-Europe, Milan, Italy, 15-19 May 2011. (with Magnus Rahmberg, Sara Nilsson)
- 6th Int. Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2011), Maribor, Slovenia 3th-7th September 2011. Platform Kovarich
- Conferentia Chemometrica 2011, Sümeg, Hungary, September 18-21, 2011, Cassani.
- 16th Int. Symposium on Environ. Pollution and its impact on Life in the Mediterranean region (MESAEP), Ioannina (Greece), 24-27 Sept. 2011. Poster.
- SETAC Italian Branch, Ecomondo (Rimini), 9 Ott. 2011, platform Cassani
- SETAC North America, 32nd Meeting, Boston (USA), 13-17 Nov. 2011.Platform Papa

GENERAL for QSAR modelling in CADASTER Project

- On the agreement of external validation parameters for linear regression QSAR models
Nicola Chirico and Paola Gramatica

- SETAC-Europe 2011 - 21th Annual Meeting SETAC-Europe, Milan, Italy, 15-19 May 2011
- 6th Int. Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resources (CMTPI-2011), Maribor, Slovenia 3th-7th September 2011.

- Linear free energy relationships and latent variable methods: Similarity in modelling environmentally relevant properties
Tomas Öberg and Tao Liu

- SETAC Europe 19th Annual Meeting in Gothenburg, May 31-June 4, 2009.

- Treatment of uncertainty from QSAR models in risk assessment
Tomas Öberg

- Annual Meeting of the Society for Risk Analysis, Baltimore, Maryland, December 6-9, 2009.

- Characterization of variability and uncertainty from QSARs for probabilistic risk assessments within REACH
Ullrika Sahlin, Monika Filipsson and Tomas Öberg

- 18th European Symposium on Quantitative Structure Activity Relationships, Rhodes, September 19-24, 2010.

- Towards guidance on how to characterize predictive uncertainty in QSAR regression models within the CADASTER project
Ullrika Sahlin, Tom Aldenberg and Tomas Öberg

- SETAC Europe 21st Annual Meeting in Milan, Italy, May 17-19, 2011.

Appendix I. QSAR/QSPR models developed for BFRs

Partner 3 UI activity

Table 1. MLR-OLS models for physico-chemical properties (Full models)

Endpoint	N_{obj}	Descriptors	R^2	Q^2_{LOO}	$AD\%_{243\ BFR}$
MP*	25	*X2A*	0.84	0.82	96%
LogVP*	34	*T(O..Br)*	0.99	0.98	83%
LogKoa*	30	*T(O..Br)*	0.97	0.97	82%
LogKow*	20	*T(O..Br)*	0.96	0.96	86%
Log H	7	*BEHe7 not Ext Val*	0.97	0.93	56%
LogWS	12	*Mor23m not Ext Val*	0.92	0.88	95%

Models were externally validated during their development (0.95 < Q2ext by different formulas < 0.99).
Papa, E.; Kovarich, S.; Gramatica P., *QSAR Comb. Sci.* (2009) 28, 790-796
Papa, E.; Kovarich, S.; Gramatica P., *Mol. Info. (*2011) 30, 232-240

Table 2. MLR-OLS models for endocrine disrupting potency (Full models).

Endpoint	N_{obj}	Descriptors	R^2	Q^2_{LOO}	$AD\%_{\ 243\ BFRs}$
Log RBA*	18	*L1v, Mor22u*	0.82	0.73	75%
Log 1/EC50ERODind	8	*piID not Ext Val*	0.85	0.75	93%
Log 1/EC50DRag	8	*Mor08e not Ext Val*	0.91	0.85	81%
Log 1/EC50ERag	8	*RGyr not Ext Val*	0.95	0.88	99%
Log 1/IC50PRant*	19	*RDF045m, GATS4m*	0.87	0.82	93%
Log T4REP*	17	*qpmax, MATS6v*	0.94	0.91	98%
Log E2SULTREP*	21	*B08[C-O], GGI7*	0.88	0.84	100%

Models were externally validated during their development (0.95 < Q2ext by different formulas < 0.99).
Papa, E.; Kovarich, S.; Gramatica, P., *Chem. Res. Toxicol.* (2010) 23, 946-954

Table 3. Classification models for endocrine disruption potency of BFRs (Full models, previously externally validated during their development).

Endpoint	N_{obj}	Descriptors	k	Sn	Sp	$AD\%_{\ 243\ BFRs}$
DR_{ag}	24	*F04[O-Br] RDF055v*	4	1	0.94	98
DR_{ant}	24	*Jhetm BEHm7*	1	1	0.87	93
ER_{ag}	24	*Ms BEHv7*	1	1	0.94	99
ER_{ant}	24	*QW nArOH*	1	1	0.94	100
AR/PR_{ant}	24	*GGI8*	1	1	1	99
$T4\text{-}TTR_{comp}$	29	*DISPe nArOH*	3	0.94	0.83	92
$E2SULT_{inh}$	29	*Mor21v qnmax*	1	0.95	1	88

Kovarich, S.; Papa, E.; Gramatica, P., *J Haz. Mat.* (2011) 190, 106-112.

Appendix II. QSAR/QSPR models developed for PFCs

Partner 3 UI activity

Table 1. MLR models for physico-chemical properties (Full models, previously externally validated during their development).

Endpoint	N_{obj}	Descriptors	R^2	Q^2_{LOO}	Q^{2*}_{EXT}	AD% $_{>200\ PFC}$
MP	94	*AAC, F02[C-F], C-013, RBF*	0.80	0.78	0.61-0.91	95%
BP	93	*ATS1m, Ms, nROH, AMW*	0.93	0.92	0.84-0.94	82%
LogVP	35	*F03[C-F], AAC,nDB*	0.91	0.88	0.80-0.88	94%
LogWS	20	*T(F..F), SIC1*	0.76	0.69	0.79-0.93	88%
LogCMC	10	*X3*	0.97	0.96	--	77%

*Range of values calculated using different Q^2_{EXT} parameters (Q^2_{EXT-F1}, Q^2_{EXT-F2}, Q^2_{EXT-F3}).

Bhhatarai, B. et al. (WP3 partners), *Molecular Informatics* (2011) 30, 189-204;

Bhhatarai, B.; Gramatica, P., *Environ. Sci. Technol.* (2010) 45 (19), 8120-8128

Table 2. MLR models for toxicity (Full models, previously externally validated during their development).

Endpoint	Nobj	Descriptors	R^2	Q^2_{LOO}	Q^{2*}_{EXT}	$AD\%_{250\ PFCs}$
Mouse Inhalation	56	*X3v; H-048; MlogP; F01[C−C]*	79.83	76.31	71.62-85.11	75.6%
Rat Inhalation	52	*Jhetv, PCR, MlogP, B02[Cl−Cl]*	78.14	73.85	66.70-75.47	76.8%
Mouse Oral	58	*HATS2u; B09[C-O]; F01[C-O]; B04[C-F]*	75.93	71.89	62.97-65.57	90.9%
Rat Oral	50	*D/Dr09; MATS1e; E1u; H8m*	88.28	85.50	80.69-91.07	83.5%

*Range of values calculated using different Q^2_{EXT} parameters (Q^2_{EXT-F1}, Q^2_{EXT-F2}, Q^2_{EXT-F3}).
Bhhatarai, B.; Gramatica, P. , *Molecular Diversity* (2011) 15, 467-476;
Bhhatarai, B.; Gramatica P., *Chem. Res. Toxicol.* (2010) 23, 528-539

Table 3. Classification models for T4-TTR competing potency (Split models for external validation).

Models	Set	N_{TR}	k	Sn	Sp	Accuracy%
AMW HATS6m	Training	10	1	1	1	100
	Prediction	9	1	1	1	100
nH HATS6m	Training	10	1	1	0.75	90
	Prediction	9	1	1	1	100
nH F06[C-O]	Training	10	1	1	1	100
	Prediction	9	1	1	0.75	90
T(F..F) HATS6m	Training	10	1	0.83	1	90
	Prediction	9	1	1	1	100

Kovarich, S., Papa, E. Gramatica P., (2011), *SAR QSAR Environ. Res.* (proceedings of CMTPI 2011), in press.

Appendix III. QSAR/QSPR models developed for Fragrances

Partner 3 UI activity

Table 1. MLR-OLS models for physico-chemical properties (Full models)

Endpoint	N_{obj}	Descriptors	R^2	Q^2_{LOO}	$AD\%_{79}$	RMSE	$RMSE_{EPISuite}$
LogKow*	52	*X2v, RDCHI*	0.82	0.79	98.73	0.47	0.64
LogWS*	37	*BEHm3, JGI3, nCconj*	0.80	0.76	97.46	0.45	0.30
LogVP*	37	*piPC01, mHDon*	0.89	0.87	100	0.2	1.91

Models were externally validated during their development (075 < Q2ext by different formulas < 0.91).

Papa, E.; Luini, M.; Gramatica, POSTER presented at SETAC-Europe 2009;

Table 2. MLR models for toxicity (Full models).

Endpoint	N_{obj}	Descriptors	R^2	Q^2_{LOO}	$AD\%_{79}$
Log 1/LD50 (mouse)	23	*H-047, nR=Cs*	0.89	0.86	97%
Log EC50 NADH-Ox (rat)	20	*nC, R5u+*	0.86	0.82	86%
Log EC50 Dψm (rat)	15	*ATS4v, MATS2m*	0.92	0.87	82%

Models were externally validated during their development (0.73<Q2ext by different formulas<0.98).

Papa, E.; Luini, M.; Gramatica, P. *SAR QSAR Environ. Res.* (2009) 20, 767–779.

Appendix IV. QSAR/QSPR models developed for B-TAZs

Partner 3 UI activity

Table 1. MLR-OLS models for physico-chemical properties (Full models, previously externally validated during their development).

Endpoint	N_{obj}	Descriptors	R^2	Q^2	Q^2_{EXT}*
LogWS	49	CIC0, AMW, MATS7e	83.81	81.15	69.19-88.14
LogKow	64	B08[C-C], nN, GATS3m, MATS1v	88.63	86.71	80.90-94.49
LogVP	33	BELp2, RBN, B09[N-Cl]	80.91	75.08	63.65-73.61
MP	56	R2e, GGI4, F03[N-N], X1A	81.32	77.34	71.93-87.58

*Range of values calculated using different Q^2_{EXT} parameters (Q^2_{EXT-F1}, Q^2_{EXT-F2}, Q^2_{EXT-F3}).
The AD to 351 (benzo)triazoles (72 in ECHA list) was verified: 89.1%-96.5%.
Bhhatarai B., Gramatica P., *Water Res.* (2011) 45, 1463-1471.

Table 2. MLR-OLS models for aquatic toxicity (Full models, previously externally validated during the model development). Papers in preparation.

OLS models developed for EC_{50} acute toxicity (**algae**, *Pseudokirchneriella subcapitata*) of (B)TAZs.

NTr.= 35 (17 (B)TAZs + 18 Other Azo-Aromatic compounds)

Model ID		Descriptors	N_{TR}	N_P	R^2	Q^2_{LOO}	Q^2_{EXT} (range)*	AD% on 386
Model 1	Split R(30)	DRAGON 5.5 (n=3)	24	11	0.85	0.79	0.73 - 0.79	
	Split K(30)		22	13	0.83	0.76	0.72 - 0.84	
	FULL		35		0.82	0.78		93.2
Model 2	Split R(30)	PaDEL-Descriptor v.2.7 (n=3)	24	11	0.83	0.75	0.70 - 0.77	
	Split K(30)		22	13	0.77	0.68	0.68 - 0.82	
	FULL		35		0.80	0.74		95.6
Model 3	Split R(30)	CADASTER online platform (n=3)	24	11	0.90	0.87	0.70 - 0.77	
	Split K(30)		22	13	0.89	0.86	0.69 - 0.82	
	FULL		35		0.85	0.81		88.9

*Range of values calculated using different Q^2_{EXT} parameters (Q^2_{EXT-F1}, Q^2_{EXT-F2}, Q^2_{EXT-F3} and CCC).
Several meeting presentations.

MLR models developed for EC_{50} acute toxicity (***Daphnia magna*****)** of (B)TAZs.

NTr.= 97 (46 (B)TAZs + 51 Other Azo-Aromatic compounds)

Model ID		**Descriptors**	**N_{TR}**	**N_P**	**R^2**	**Q^2_{LOO}**	**Q^2_{EXT} (range)***	**AD% On 386**
Model 1	Split R(30)	DRAGON 5.5 (n=5)	65	32	0.78	0.74	0.69 - 0.71	
	Split K(30)		65	32	0.75	0.70	0.78 – 0.83	
	FULL		97		0.77	0.74		90.7
Model 2	Split R(30)	CADASTER online platform (n=3)	65	32	0.74	0.71	0.71 - 0.73	
	Split K(30)		65	32	0.74	0.70	0.73 - 0.79	
	FULL		97		0.75	0.72		89.9

*Range of values calculated using different Q^2_{EXT} parameters (Q^2_{EXT-F1}, Q^2_{EXT-F2}, Q^2_{EXT-F3} and CCC).
Several meeting presentations.

MLR models developed for LC_{50}-96h (fish, ***Oncorhynchus mykiss***) of (B)TAZs.

NTr.= 76 (28 (B)TAZs + 48 Other Azo-Aromatic compounds)

EV1=10 (B)TAZs
EV2= 8 (B)TAZs

Model ID		**Descriptors**	**N_{TR}**	**N_P**	**R^2**	**Q^2_{LOO}**	**Q^2_{EXT} (range)***	**AD% On 386**
Model 1	Split R(30)	DRAGON 5.5 (n=4)	53	23	0.80	0.76	0.86 - 0.90	
	Split K(30)		53	23	0.82	0.79	0.79 - 0.87	
	FULL		76		0.82	0.79		93.7
	EV1		10		0.82	0.79	0.85-0.87	
	EV2		08		0.82	0.79	0.79-0.89	
Model 2	Split R(30)	PaDEL-Descriptor v.2.7 (n=4)	53	23	0.81	0.77	0.72-0.80	
	Split K(30)		53	23	0.81	0.77	0.71-0.82	
	FULL		76		0.79	0.76		95.5
	EV1		10		0.79	0.76	0.75-0.79	
	EV2		08		0.79	0.76	0.72-0.86	

*Range of values calculated using different Q^2_{EXT} parameters (Q^2_{EXT-F1}, Q^2_{EXT-F2}, Q^2_{EXT-F3} and CCC).
Several meeting presentations.

Comparison of UI models with EPI Suite (ECOSAR) for TAZs aquatic toxicities

ALGAE

MODEL	RMSE(12TAZs)
DRAGON	0.34
PaDEL	0.29
CADASTER	0.41
CONSENSUS	0.31
EPI (BASELINE eq)	0.81
EPI (TAZs eq)	0.51

DAPHNIA

MODEL	RMSE(32TAZs)
DRAGON	0.44
CADASTER	0.45
EPI (BASELINE eq)	0.67
EPI (TAZs eq)	0.63

FISH

MODEL	RMSE (33TAZs)
ECOSAR(Triazole not fused)	0.839
DRAGON	0.473
PADEL	0.533
Consensus	0.474

MODEL	RMSE (46 BTAZs)
ECOSAR (baseline)	0.959
DRAGON	0.454
PADEL	0.537
Consensus	0.456

Partner 6 HMGU activity

The analysed descriptor sets included:

ADRIANA.Code (3D) comprises 211 molecular descriptors based a sound geometric and physicochemical basis. The classes of descriptors cover global molecular descriptors, shape and size descriptors, topological and 3D property-weighted autocorrelation descriptors.[4]
CDK (3D) included topological, geometrical, constitutional, electronic and hybrid descriptors.[5] In total 204 descriptors were calculated.
ChemAxon descriptors (3D) included elemental analysis, charge, geometry, partitioning, protonation, isomers and "other" descriptors.[6]
Dragon6 (3D) represented the largest pool, which included 4885 descriptors grouped in 29 different blocks.[7]
E-state indices[8,9] (2D) were calculated using E-state program, which was used to predict logP and water solubility in the ALOGPS program.[10] The logP and logS values calculated using ALOGPS 2.1 version were also included.
ISIDA Fragmentor (2D)[11] was used to calculate augmented atoms of length 3 to 5.
GSFRAG (2D) included descriptors based on fragments that contain a labeled vertex, allowing one to capture the effect of heteroatoms.[12]
Inductive descriptors (3D), which are based on LFER (Linear Free Energy Relationships) equations for inductive and steric substituent constants, were implemented according to ref [13].
Mera (3D) included geometrical, energy characteristics and physicochemical descriptors[14]. In this set we also included MERSY, which estimates molecular symmetry and chirality.
Shape Signatures (3D) encoded spatial shape characteristics of molecules using ray tracing, which explores volume enclosed by the solvent accessible surface of a molecule [15].
Spectrophores fingerprints (3D) are calculated as one-dimensional compression of molecular properties fields surrounding molecules.

1. Tetko, I. V., Associative neural network. *Neural Process. Lett.* **2002,** *16* (2), 187-199.
2. Tetko, I. V., Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (3), 717-728.
3. Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, A. N.; Zefirov, N. S., Fragmental descriptors with labeled atoms and their application in QSAR/QSPR studies. *Doklady Chemistry* **2007,** *417*, 282-284.
4. Gasteiger, J., Of molecules and humans. *J. Med. Chem.* **2006,** *49* (22), 6429-34.
5. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E., The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (2), 493-500.
6. Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, II; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V., Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput.-Aided Mol. Des.* **2011,** *25* (6), 533-54.
7. Todeschini, R.; Consonni, V., *Handbook of Molecular Descriptors*. WILEY-VCH: Weinheim, 2000; p 667.
8. Kier, L. B.; Hall, L. H., *Molecular Structure Description: The Electrotopological State*. Academic Press: London, 1999; p 245.
9. Hall, L. H.; Kier, L. B., Electrotopological state indices for atom types - a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995,** *35* (6), 1039-1045.
10. Tetko, I. V.; Tanchuk, V. Y., Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.* **2002,** *42* (5), 1136-1145.
11. Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G., ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008,** *4* (3), 191-198.
12. Stankevich, I. V.; Skvortsova, M. I.; Baskin, I. I.; Skvortsov, L. A.; Palyulin, V. A.; Zefirov, N. S., Chemical graphs and their basis invariants. *Journal Of Molecular Structure-Theochem* **1999,** *466*, 211-217.
13. Cherkasov, A.; Jonsson, M., Substituent effects on thermochemical properties of free radicals. New substituent scales for C-centered radicals. *J. Chem. Inf. Comput. Sci.* **1998,** *38* (6), 1151-1156.
14. Potemkin, V. A.; Grishina, M. A.; Bartashevich, E. V., Modeling of drug molecule orientation within a receptor cavity in the BiS algorithm framework. *Journal of Structural Chemistry* **2007,** *48* (1), 155-160.

15. Zauhar, R. J.; Moyna, G.; Tian, L.; Li, Z.; Welsh, W. J., Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. *J. Med. Chem.* **2003,** *46* (26), 5674-90.

Partner 7 IdeaConsult activity.

MLR models developed for LC_{50}-96h (fish, *Oncorhynchus mykiss*) of (B)TAZs.

NTr.= 76 (28 (B)TAZs + 48 Other Azo-Aromatic compounds)

Test set EV1=10 (B)TAZs
Validation set EV2= 8 (B)TAZs

A list of descriptors selected in the final models follows:

Mp mean atomic polarizability (scaled on Carbon atom)
nN number of Nitrogen atoms
SIC1 structural information content (neighborhood symmetry of 1-order)
EEig07d Eigenvalue 07 from edge adj. matrix weighted by dipole moments
C-024 Atom-centred fragment: R--CH--R
O-058 Atom-centred fragment: =O
nBzn number of benzene-like rings
nCIC number of rings
LP1 Lovasz-Pelikan index (leading eigenvalue)
X0Av average valence connectivity index chi-0
PW3 path/walk 3 - Randic shape index
nSO2N number of sulfonamides / sulfinamides / sulfenamides (thio- / dithio-)
VEA1 eigenvector coefficient sum from adjacency matrix
PW5 path/walk 5 - Randic shape index
GATS2e Geary autocorrelation - lag 2 / weighted by atomic Sanderson electrone-gativities
BEHm2 highest eigenvalue n. 2 of Burden matrix / weighted by atomic masses
BELm2 lowest eigenvalue n. 2 of Burden matrix / weighted by atomic masses
JGI3 mean topological charge index of order3
nHDon number of donor atoms for H-bonds (N and O)
X3A average connectivity index chi-3
C-001 Atom-centred fragment: CH3R / CH4
RBF rotatable bond fraction
nR05 number of 5-membered rings

The following statistics have been calculated:
R2 Coefficient of determination (R2 = 1 –RSS/TSS)
Q2 Cross-validated R2 - leave-one-out
Q2boot boot strap coefficient of prediction
Q2ext Coefficient of prediction for the external test set (explained variance in prediction)
a(R2) Y-scrambling parameter for the learning set
a(Q2) Y-scrambling parameter for the testing set
R2 adjusted adjusted coefficient of determination
LOF Lack of Fit /Friedman modified/
AIC Akaike information
Kx total correlation in the model predictors
KXY total correlation in the set given by the model predictors X plus the response Y
SDEP Standard Deviation Error in Prediction
SDEC Standard Deviation Error in Calculation
F Fisher function
s residual standard deviation
DF degree of freedom
DK Quick rule threshold

DQ Q asymptotic rule threshold
Rp redundancy rule threshold
Rn overfitting rule threshold
TSS Total Sum of Squares
AVH average leverage value
Ax1, Ax2 the two model distance coordinates
Pop the population the model belongs to

The set of models with automatically selected variables (Models A) and with expert selection (Models 1-3) are compared with the purpose of identifying the best models (Fig.1). As expected, the increase of model complexity (number of variables) improves model accuracy, but decreases the models performance on test datasets (Models A on set EV1, EV2). The performance of models 1-3 on training, test and validation set is relatively uniform, and an indication of better stability.

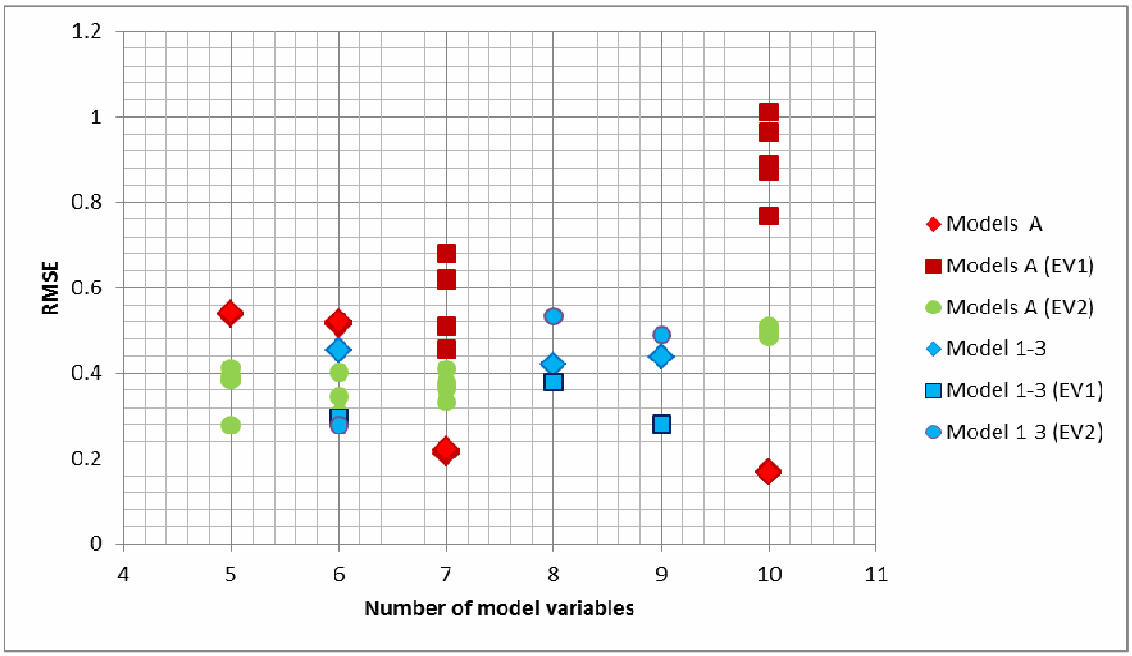

Fig. 1

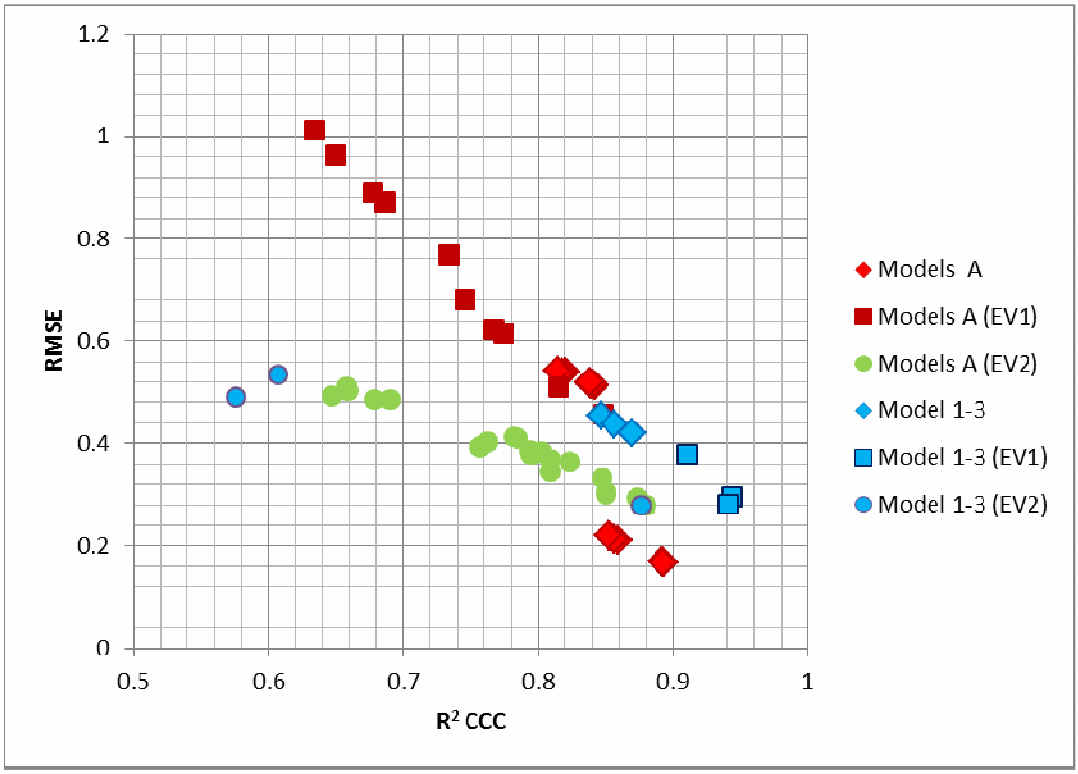

Fig. 2

We report regression equations and performance statistics on Models 1-3. Details on the rest of the models are available as separate files.

Model 1. Linear regression, 6 variables

Descriptor	Description	Regression coefficient	Relative standard deviation,%	Standard deviation
	Intercept	2.671	31.6	0.844
G_(P)	Number of P atoms	1.263	27.9	0.353
Mp	Dragon	10.561	13.0	1.374
nN	Dragon	-0.135	30.5	0.041
SIC1	Dragon	-7.742	8.8	0.682
EEig07d	Dragon	0.598	12.3	0.074
O-058	Dragon	-0.253	20.4	0.052

Model 2 . Linear regression, 9 variables

Descriptor	Description	Regression coefficient	Relative standard deviation,%	Standard deviation
	Intercept	5.741	7.2	0.415
G_(P)	Number of P atoms	1.145	33.6	0.384
G_(S)	Number of S atoms	-0.43	33.9	0.146
ln(Mp)	Dragon	8.758	11.2	0.978
nCIC	Dragon	-0.346	30.1	0.104
ln(SIC1)	Dragon, ln()	-6.218	7.3	0.457
EEig07d	Dragon	0.609	14.1	0.086
$O\text{-}058^2$	Dragon, power(2)	0.028	49.2	0.014

$nHDon^2$	Dragon, power(2)	-0.034	28.7	0.013
$nR05^2$	Dragon, power(2)	0.169	31.9	0.054

Model A5 & A6 (Automatic selection of variables)

	Linear regression, 5 variables			Linear regression, 6 variables		
Descriptor	Description	Regression coefficient	Conf.Intervals (.95)	Descriptor	Regression coefficient	Conf.Intervals (.95)
	Intercept	2.945	1.813	Intercept	1.262	2.081
Mp	Dragon	10.805	2.957	Mp	8,750	3.154
nN	Dragon	-0.132	0.089	nN	-0.143	0.085
SIC1	Dragon	-8.296	1.432	SIC1	-7.074	1.604
EEig07d	Dragon	0.582	0.159	EEig07d	0.599	0.152
O-058	Dragon	-0.259	0.111	O-058	-0.246	0.106
				X0Av	3.711	2.560

Statistics.

Models	**Model 1**			**Model 2**			**Model 3**		
Statistics\ Data sets	**Training set**	**EV1**	**EV2**	**Training set**	**EV1**	**EV2**	**Training set**	**EV1**	**EV2**
RMSE	0.454	0.296	0.279	0.438	0.282	0.490	0.419	0.372	0.534
R^2 classical [a]	0.852	0.954	0.880	0.862	0.946	0.796	0.874	0.920	0.738
R^2 (1 - RSS/TSS)	0.852			0.862			0.874		
CCC[b]	0.846	0.943	0.876	0.857	0.941	0.576	0.870	0.910	0.607
Fisher function	58.259			42.437			53.065		
Residual standard deviation	0.479			0.472			0.448		
LOO RMSE	0.546			0.550			0.542		
LOO R^2 classical *	0.787			0.785			0.791		
LOO Q^2	0.785			0.782			0.788		
LOO CCC[b]	0.787				0.785			0.791	
Y scrambling R^2_YS(average)	0.0759			0.1134			0.1031		
Q^2(F1) [b,c]		0.939	0.905		0.945	0.706		0.904	0.650
Q^2(F2) [b,c]		0.938	0.880		0.944	0.629		0.902	0.559
Q^2(F3) [b,c]		0.938	0.944		0.943	0.827		0.901	0.794

a. R^2 calculated by the classical formula of Pearson product moment correlation coefficient

b. Calculated, as defined *in Nicola Chirico and Paola Gramatica,* 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.*, 2011, *51* (9), pp 2320–2335

c. Calculated, as defined in *Viviana Consonni, Davide Ballabio, and Roberto Todeschini, Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 2009, 49, 1669–1678*