Development of models according to the OECD principles

Ester Papa,
Paola Gramatica

QSAR Research Unit in Environmental Chemistry and Ecotoxicology
DBSF -University of Insubria, Varese - Italy
ester.papa@uninsubria.it
http://www.qsar.it

SEVENTH FRAMEWORK PROGRAMME

OUTLINE

1) QSAR in Regulation - OECD Principles

2) Modelling strategy

3) Examples: CADASTER Models

QSAR in Regulation

Increasing interest in the development and validation of alternative methods, in vitro and in silico, such as QSARs, to minimize costs and animal lives

In silico predictions can be used to:

- **highlight chemicals (more/less hazardous, alternatives..)**
- **prioritize chemicals and focus experimental tests**
- **fill data gaps (ITS applications)**

QSAR in Regulation

- The REACH REGULATION (1907/2006/EC)
- The new COSMETIC DIRECTIVE (76/768/EEC)
- The new BIOCIDE REGULATION (EU) No 528/2012

Acceptability of QSARs in Regulation

- Regulatory need
- Free public availability
- Transparency
- Communication

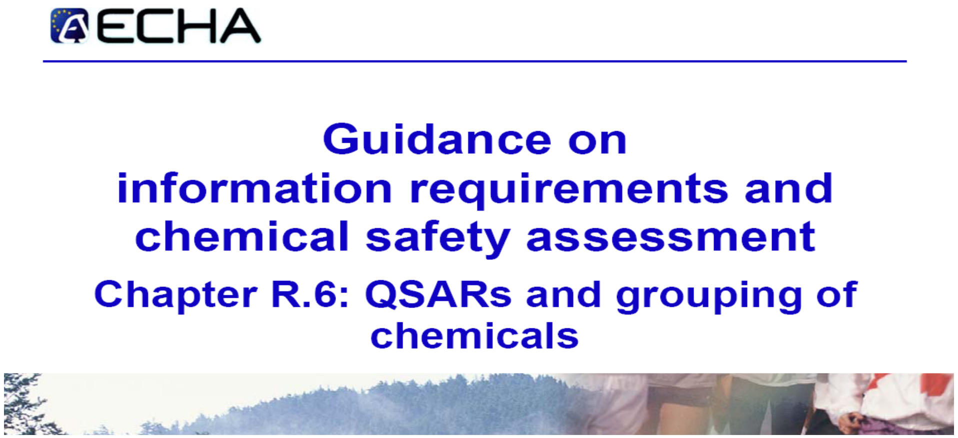
OECD Principles for QSAR models (2004)

1. a defined endpoint
2. an unambiguous algorithm
3. a defined domain of applicability
4. appropriate measures of goodness of fit, robustness and predictivity
5. a mechanistic interpretation, if possible

REACH and ECHA Guidance

QSAR can be used, instead of tests, depending on:

1. Scientific validity of the model (i.e. OECD Principles)
2. Inclusion in the model domain
3. Adequacy of the endpoint to the regulatory context



ECHA

Guidance on
information requirements and
chemical safety assessment

Chapter R.6: QSARs and grouping of
chemicals

- to establish validity, and adequacy of (Q)SAR models
- to document the regulatory use of (Q)SAR models

QMRF and JRC QSAR Model Database

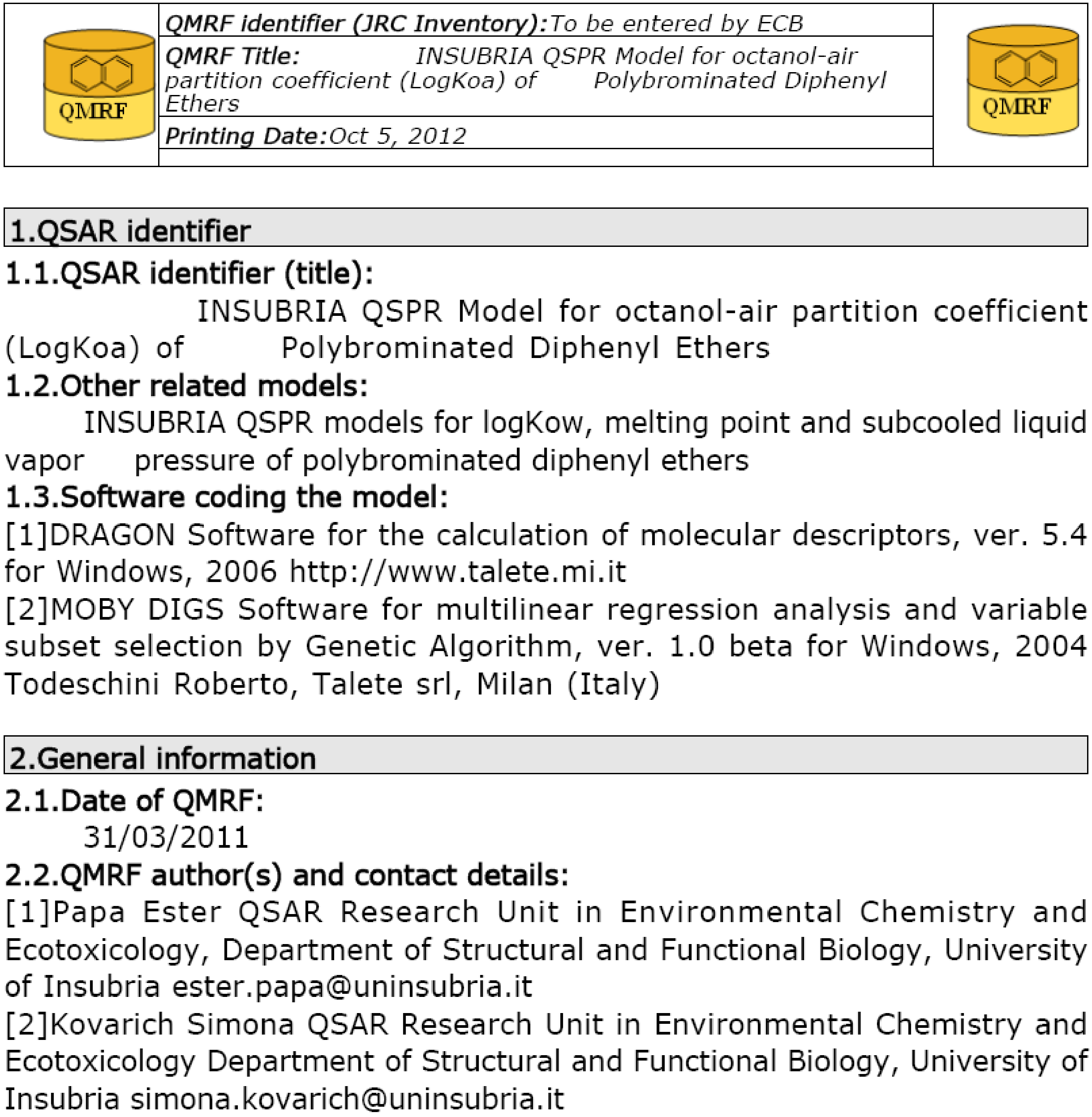
The QSAR Model Reporting Format (QMRF):

- harmonised template for summarising and reporting key information on (Q)SAR models
- structured according to the OECD (Q)SAR validation principles
- includes the results of any validation studies
- freely accessible

The QMRF is expected to be a communication tool between industry and the authorities under REACH.

JRC - QSAR Model Database is a freely accessible repository of QMRF

QMRF



QMRF identifier (JRC Inventory):To be entered by ECB
QMRF Title: INSUBRIA QSPR Model for octanol-air partition coefficient (LogKoa) of Polybrominated Diphenyl Ethers
Printing Date:Oct 5, 2012

1.QSAR identifier

1.1.QSAR identifier (title):

INSUBRIA QSPR Model for octanol-air partition coefficient (LogKoa) of Polybrominated Diphenyl Ethers

1.2.Other related models:

INSUBRIA QSPR models for logKow, melting point and subcooled liquid vapor pressure of polybrominated diphenyl ethers

1.3.Software coding the model:

[1]DRAGON Software for the calculation of molecular descriptors, ver. 5.4 for Windows, 2006 http://www.talete.mi.it

[2]MOBY DIGS Software for multilinear regression analysis and variable subset selection by Genetic Algorithm, ver. 1.0 beta for Windows, 2004 Todeschini Roberto, Talete srl, Milan (Italy)

2.General information

2.1.Date of QMRF:

31/03/2011

2.2.QMRF author(s) and contact details:

[1]Papa Ester QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria ester.papa@uninsubria.it

[2]Kovarich Simona QSAR Research Unit in Environmental Chemistry and Ecotoxicology Department of Structural and Functional Biology, University of Insubria simona.kovarich@uninsubria.it

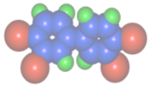
QSARs based on the OECD principles

1. Defined end-points: LogKow, Rodents toxicity
2. Unambiguous algorithm.

✓ Chemical representation by theoretical molecular descriptors (DRAGON)

✓ Statistical method → MLR regression (OLS); variable selection by Genetic Algorithms (GA)

3. Applicability Domain: → leverage approach (MLR) / graphic analysis
4. Validation for model stability and predictivity (internal and external validation)
5. Interpretation of molecular descriptors



Unambiguous algorithm

Chemical representation by theoretical molecular descriptors

- Calculated from the chemical structure.
- Different types of molecular representation: different "views" on a molecule. (This is necessary to perform structural similarity studies)
- Higher possibility to catch structural features related to the studied end point. (No *a priori* bias on hypothesized mechanism).

MLR regression (OLS)

- Reduce complexity (Ockham's Razor)

Variable reduction and selection

- **Variable Reduction**

- **Variable Selection by Genetic Algorithm (GA)**

Optimisation Parameters for GA in MLR

Q2 (LOO) *leave-one-out* by applying the QUIK rule (KXY-KXX = ΔK should be > 0)

Models with higher ΔK, among models with similar Q^2 (LOO), are then checked by a stronger validation

Applicability Domain by Leverage

MLR

$$\hat{y} = X(X^TX)^{-1}X^T y = \underline{H}\ y$$

The *i^th^* main diagonal entry of $\underline{H}$ (the **Hat matrix**) (h_{ii}) provides a measure of how far observation *i* is from the center of the X data (leverage)

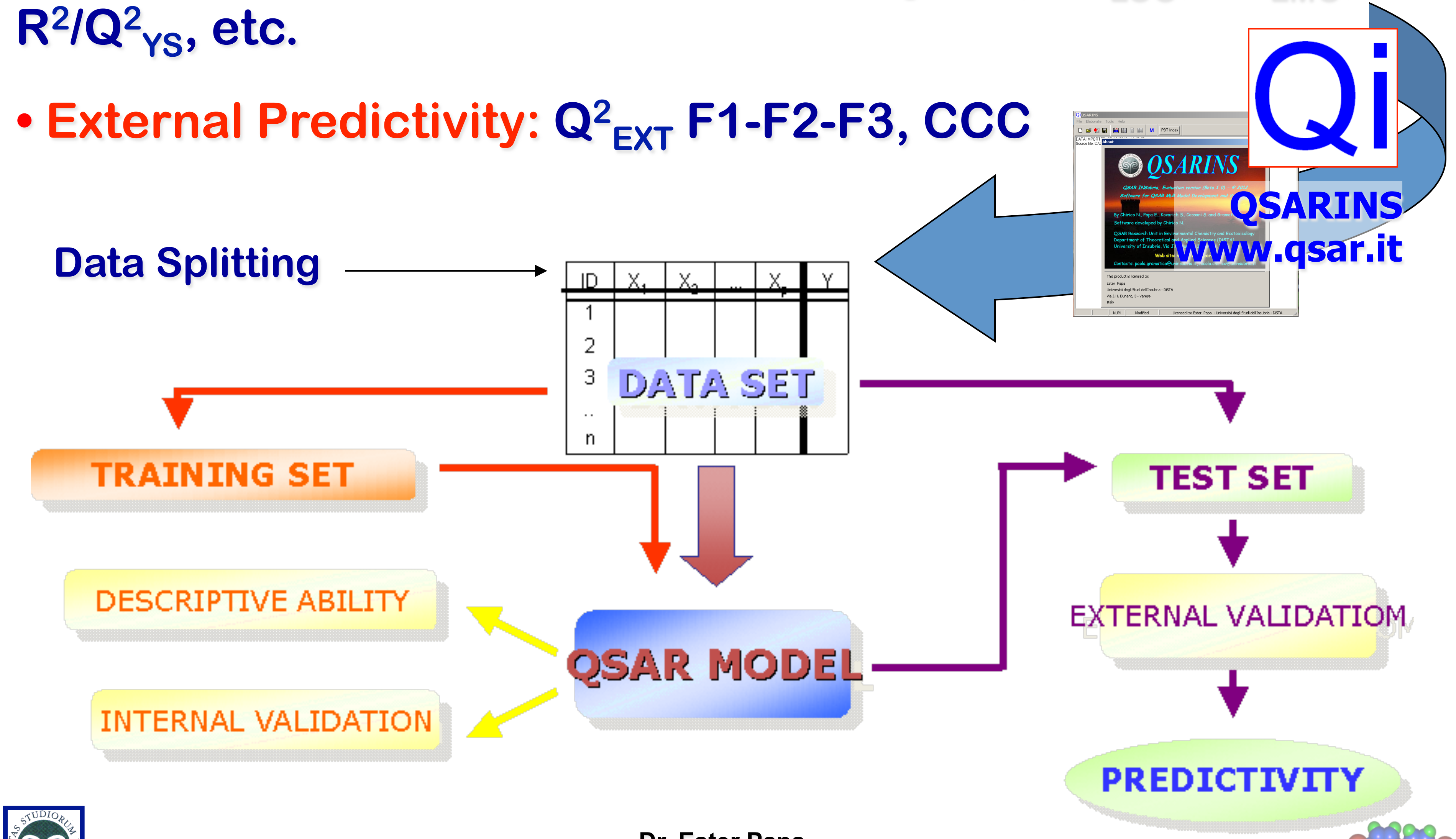
Cut off value = $h^* = 3(p+1)/n$

A chemical with a HIGH LEVERAGE is STRUCTURALLY ANOMALOUS in the CHEMICAL DOMAIN of the model:

- **in the TRAINING: influences the regression (selection of descriptors and of MLR parameters).**
- **in the TEST: predictions are extrapolated, less reliable.**

Evaluation of the predictivity

- Internal Robustness and Predictivity: R^2, Q^2_{LOO}, Q^2_{LMO}, R^2/Q^2_{YS}, etc.
- External Predictivity: Q^2_{EXT} F1-F2-F3, CCC



QSARINS

DATA IMPORTE…
Source file: C:\…

QSARINS

QSAR INSubria, Evaluation version (Beta 1.0) - © 2012
Software for QSAR MLR Model Development and V…

By Chirico N., Papa E., Kovarich S., Cassani S. and Gramati…
Software developed by Chirico N.

QSAR Research Unit in Environmental Chemistry and Ecotoxicology
Department of Theoretical and Applied Sciences (DiSTA)
University of Insubria, Via J.H. Dunant, 3 …

Web site: h…sar…

Contacts: paola.gramatica@uninsubria.it, nicola.chirico@uninsubria.it

This product is licensed to:
Ester Papa
Università degli Studi dell'Insubria - DiSTA
Via J.H. Dunant, 3 - Varese
Italy

Licensed to: Ester Papa - Università degli Studi dell'Insubria - DiSTA

Dr. Ester Papa
QSAR Research Unit - DiSTA - University of Insubria - Varese (Italy)

Interpretation of descriptors

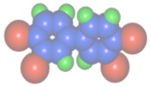
An endpoint can be the result of a series of complex mechanisms, which often can't be modeled by easily interpretable descriptors, a priori selected by the modeler

Descriptive QSAR

- Local models
- Fitting ability (high R2)
- Mechanistic interpretation of descriptors: relevant
- Application: mechanism understanding, chemical (drug) design

Predictive QSAR

- Global models
- Rigorous Validation: Internal and External Predictivity
- Interpretation of descriptors: if possible
- Application: screening/prioritization of chemicals

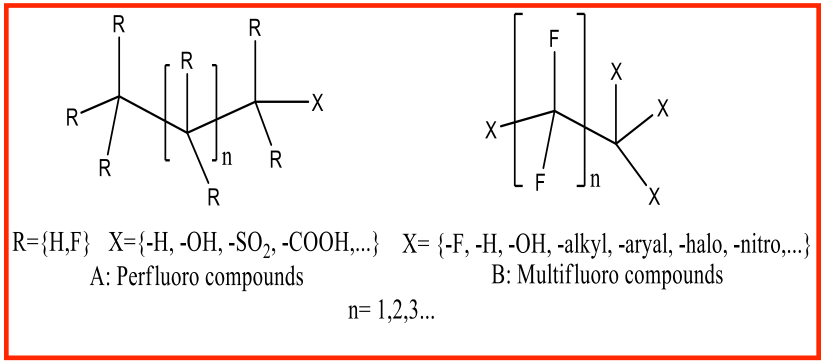


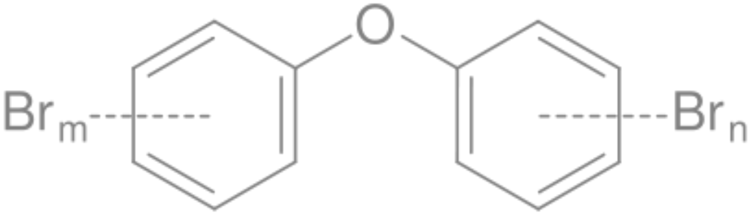
Examples

Development of models according to the OECD Principles

The FP7 Project CADASTER

Problems for PFCs and PBDEs in CADASTER


R
R
R
R
X
n
R
R
R
F
X
X
X
F
n
X
R={H,F} X={-H, -OH, -SO2, -COOH,...}
A: Perfluoro compounds
X= {-F, -H, -OH, -alkyl, -aryal, -halo, -nitro,...}
B: Multifluoro compounds
n= 1,2,3...



Limited ecotoxicological data have been found and not in reasonable amount to develop QSAR models on the endpoint of interest (i.e. SIDS)

Existing QSAR models are not always reliably applicable to PFCs and PBDEs: they are mainly out of the AD

Use of small Datasets **Use of non SIDS endpoints**

l Dataset	n° of available exp.data (→modelled)	Bibliography	Comparison with other *ad hoc* models
Henry Low Constant (H) (Pa m3/mol, 25°C)	12 → 7	Cetin & Odabasi (2005) Tittlemeier et al. (2002)	Xu et al. (2007)
Melting Point (T_M °C)	26	Kuramochi et al. (2007) Tittlemeier et al. (2002) Palm et al. (2002) Marsh et al. (1999)	not available
Vapour Pressure (Pv) (Pa, 25°C)	39 → 35	Wania & Dungani (2003) Tittlemeier et al. (2002) Palm et al. (2002) Wong et al. (2001)	Xu et al. (2007)
Water Solubility (S) (mol/L, 25°C)	13 → 12	Kuramochi et al. (2007) Wania & Dungani (2003) Tittlemeier et al. (2002) Palm et al. (2002)	not available
Log Koa	30	Gouin and Harner (2003) Harner & Shoeib (2002) Wania et al. (2002)	Xu et al. (2007) Chen et al. (2003)
Log Kow	20	Kuramochi et al. (2007) Wania & Dungani (2003) Braekevelt et al. (2003) Palm et al. (2002)	not available
Log K photolysis	15	Eriksson et al. (2004)	Niu et al. (2006) Chen et al. (2007)
Log HL photolysis	15	Eriksson et al. (2004)	not available
Log K hydrolysis	7	Rahm et al. (2005)	not available
Log HL hydrolysis	7	Rahm et al. (2005)	not available

Models

Endpoint	Obj. training	Descriptors	R^2%	Q^2%	$Q^2_{\text{EXT (rand50\%)}}$%	AD% (209 PBDEs)
logH	7	BEHe7	96.87	93.34		64.7
MP	26	X2A	84.56	82.24	88.55	97.61
$\log P_L$	34	T(O…Br)	98.63	98.45	98.62	91.38
$\log W_{sol}$	12	Mor23m	91.8	88.55		95.69
LogKoa	30	T(O…Br)	97.37	96.78	95.17	92.34
LogKow	20	T(O…Br)	96.44	95.63	91.6	96.65
$\text{Log}\,k_{\text{photol.}}$	15	MW	94.91	93.83		92.82
$\text{Log}\,k_{\text{hydrol.}}$	7	HATS2p	91.19	85.05		73.68
$\text{Half-Life}_{\text{photol.}}$	15	T(O…Br)	94.39	92.66		86.6
$\text{Half-Life}_{\text{hydrol.}}$	7	PW3	96.22	92.07		88.99

Focus on some aspects of interest:

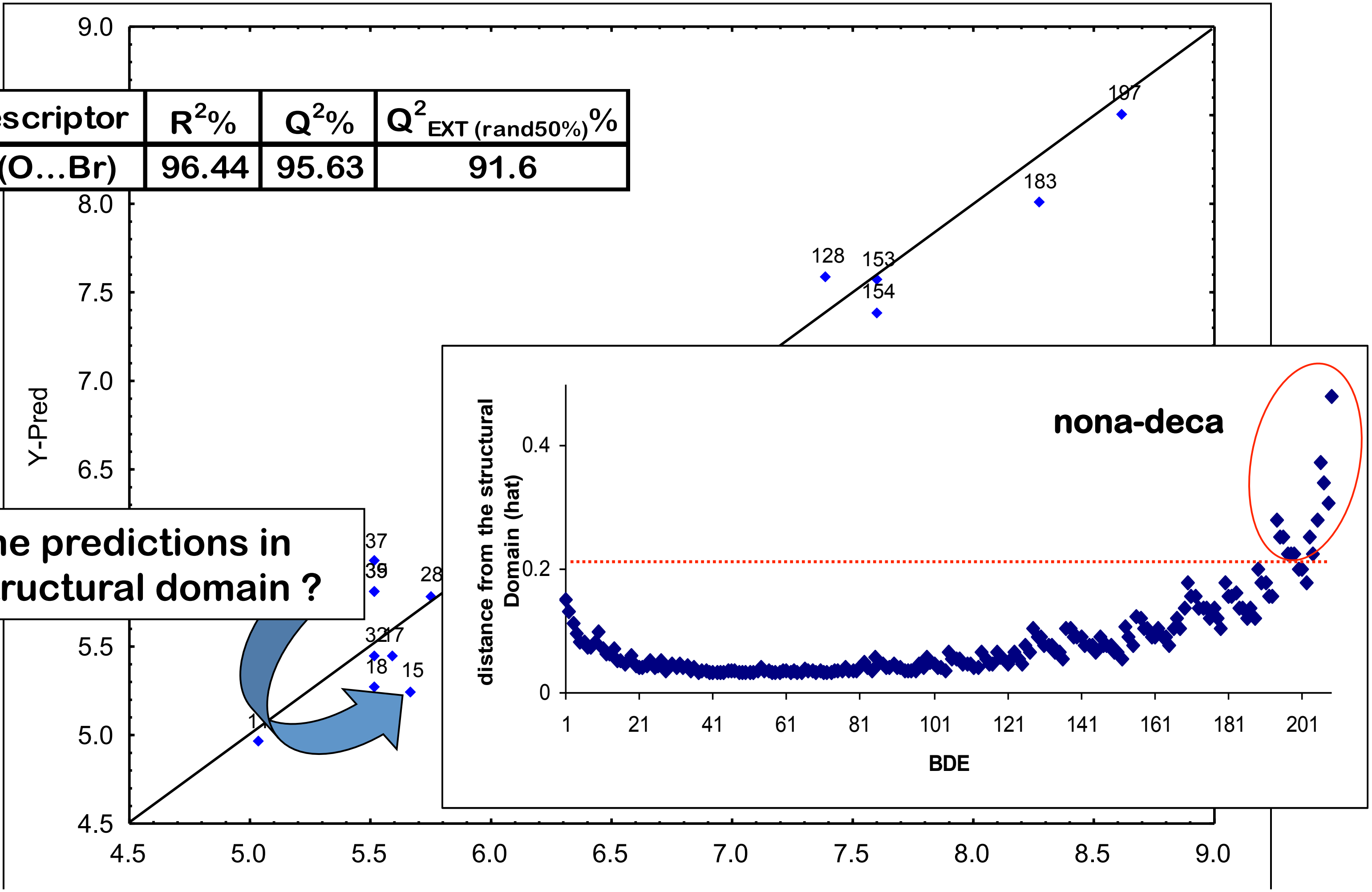
VALIDATION, DOMAIN, COMPARISON

Papa E. et al. QSAR and Combinatorial Science, 2009, 28, 790-796.

Model for Log Kow

LogKow= 3.675 + 0.162 T(O...Br)

n° Obj.	Descriptor	R^2%	Q^2%	$Q^2_{EXT\ (rand50\%)}$%
20	T(O…Br)	96.44	95.63	91.6



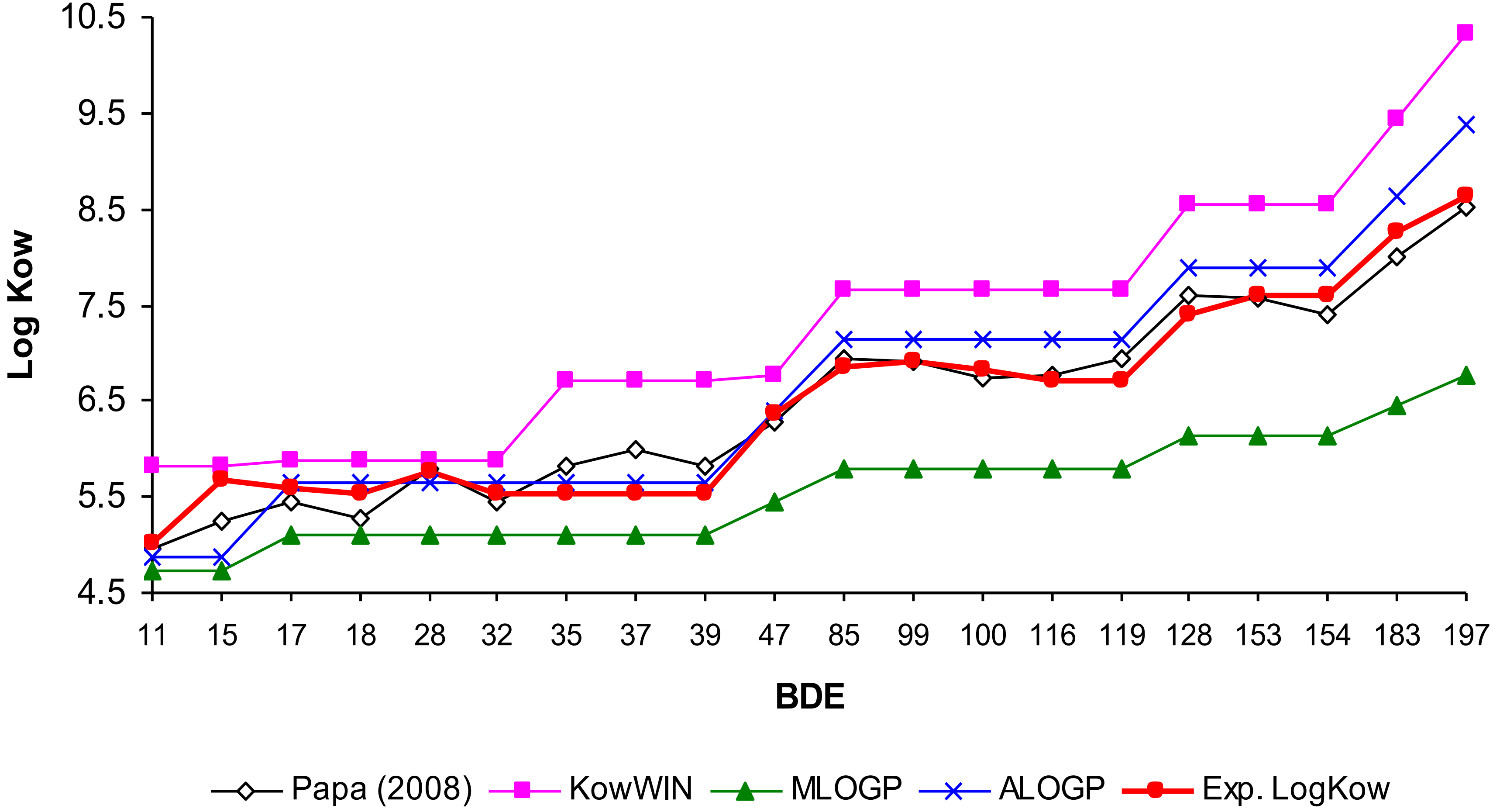
Are the predictions in the structural domain ?

Experimental range of LogKow: 5.03 (di-BDE) – 8.62 (octa-BDE)

Comparison with other calculated LogKow

Predicted and Experimental data for 20 PBDEs

Papa E. et al. Molecular Informatics 2011, 30, 232–240



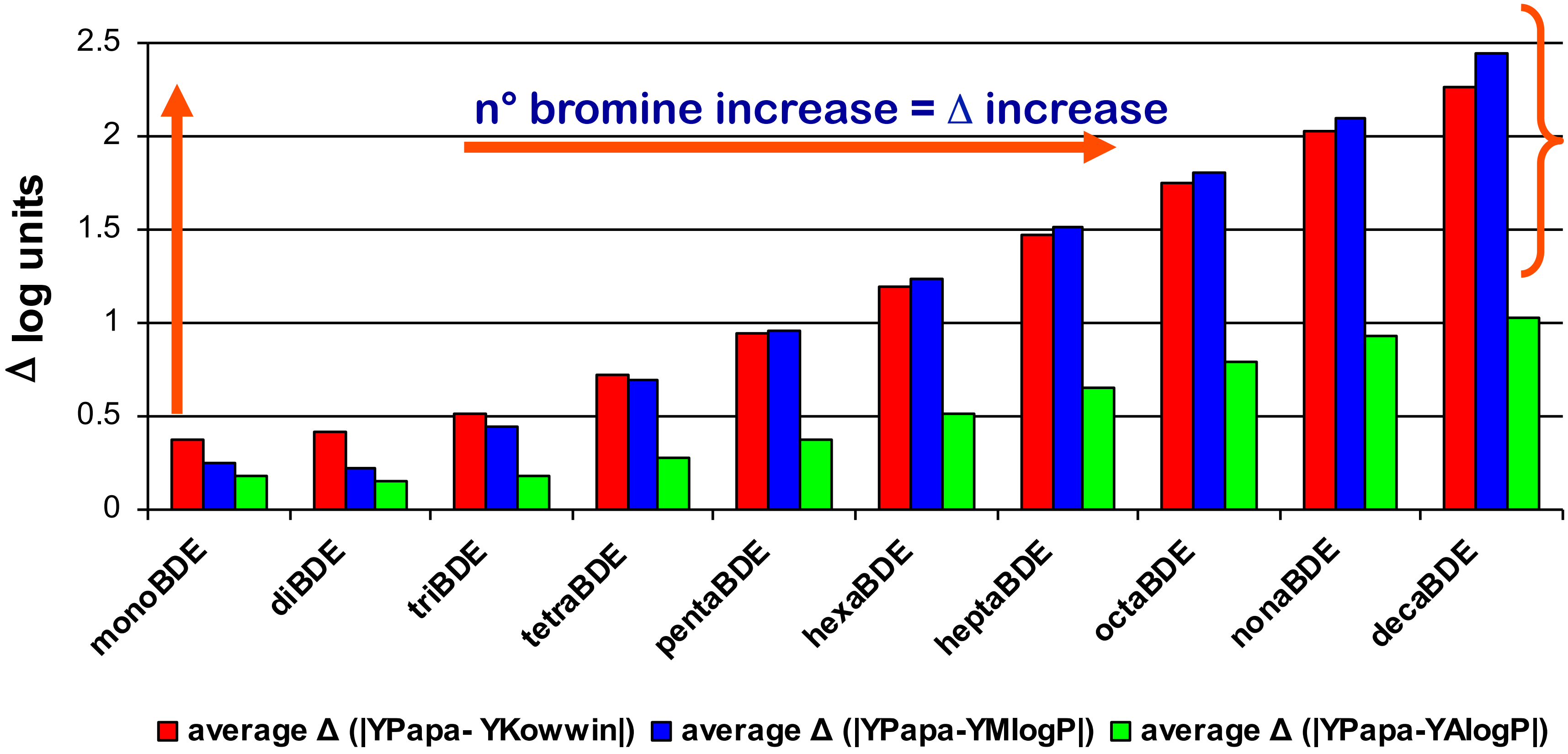
Experimental range of LogKow: 5.03 (di-BDE) – 8.62 (octa-BDE)

Dr. Ester Papa
QSAR Research Unit - DiSTA - University of Insubria - Varese (Italy)

Comparison with other calculated LogKow

Predictions for 209 PBDEs

Papa E. et al. Molecular Informatics, 2011, 30, 232–240.



Y_{Papa} = Pred. by our model (range of LogKow: 4.2 – 9.8)

Y_{Kowwin} = Pred. by Kowwin (Δmax = 2.27 log units; range of LogKow: 4.1 – 12.1)

Y_{MlogP} = Pred. by MLogP (Δmax =2.45 log units; range of LogKow: 4.1-7.4)

Y_{AlogP} = Pred. by ALogP (Δmax = 1.15 log units; range of LogKow: 4.1 –10.9)

PFCs toxicity: performances of the models

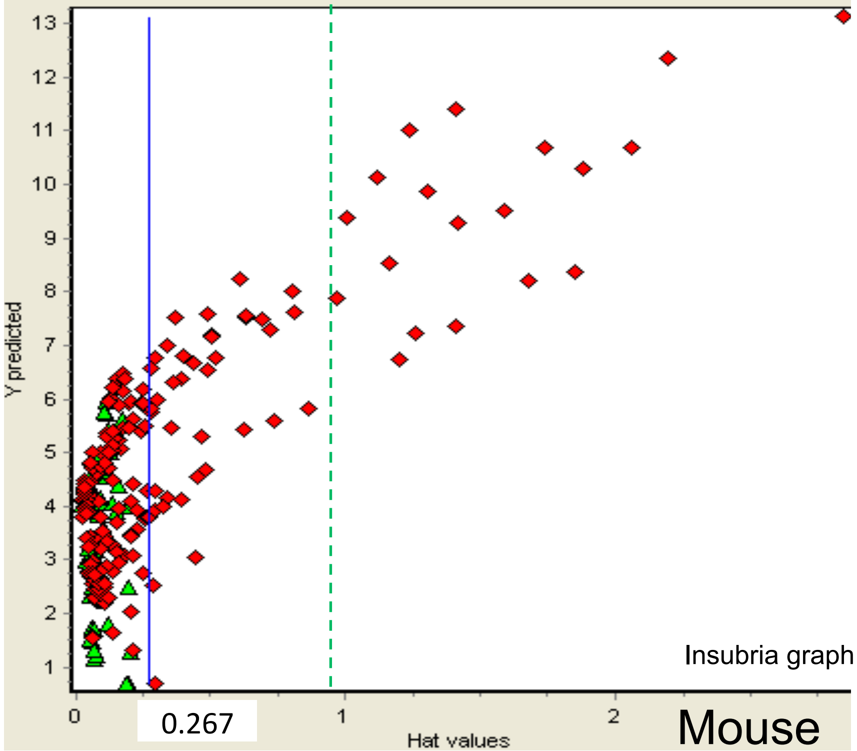
Endpoint	Descriptors	N_{obj}	R^2	Q^2_{LOO}	Q^2_{EXT}	$RMSE_{CV}$	$AD\%_{250\ PFCs}$
Mouse Inhalation	X3v; H-048; *MlogP*; F01[C−C]	*56*	*79.8*	*76.3*	*71.6-85.1*	*0.74*	*75.6%*
Rat Inhalation	Jhetv, PCR, *MlogP,* B02[Cl−Cl]	*52*	*78.1*	*73.9*	*66.7-75.5*	*0.86*	*76.8%*

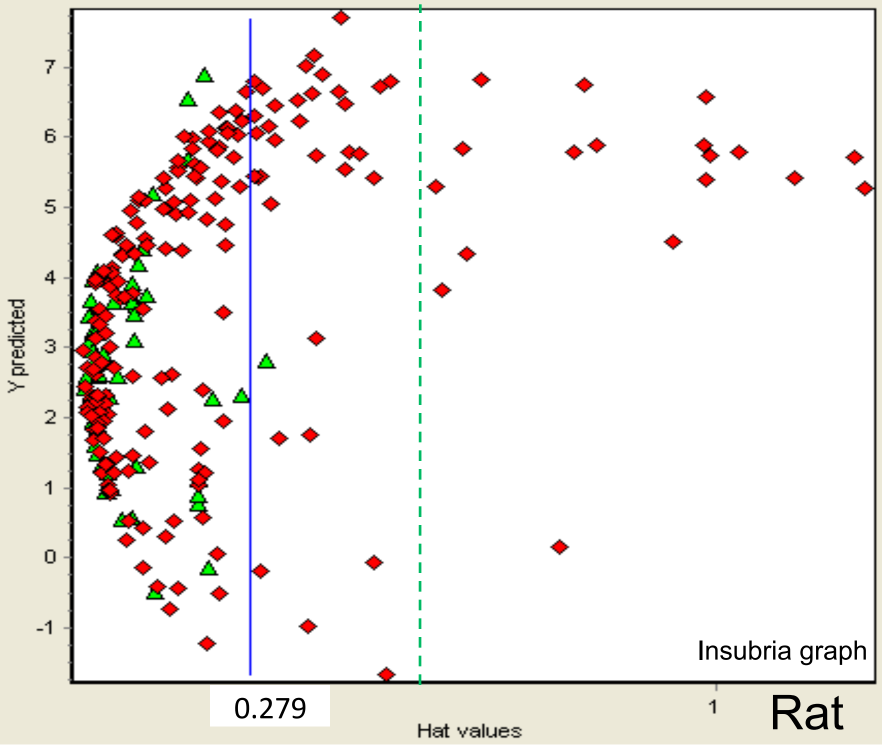
Endpoint	Descriptors	N_{obj}	R^2	Q^2_{LOO}	Q^2_{EXT}	$RMSE_{CV}$	$AD\%_{376\ PFCs}$
Mouse Oral	HATS2u; B09[C-O]; F01[C-O]; B04[C-F]	*58*	*75.9*	*71.9*	*63.0-65.*	*0.42*	*90.9%*
Rat Oral	D/Dr09; MATS1e; E1u; H8m	*50*	*88.3*	*85.5*	*80.7-91.1*	*0.47*	*83.5%*

Bhhatarai, B.; Gramatica P., Chem. Res. Toxicol., 2010, 23, 528-539.

Bhhatarai, B.; Gramatica, P.,2011, 15 (2), 467-476

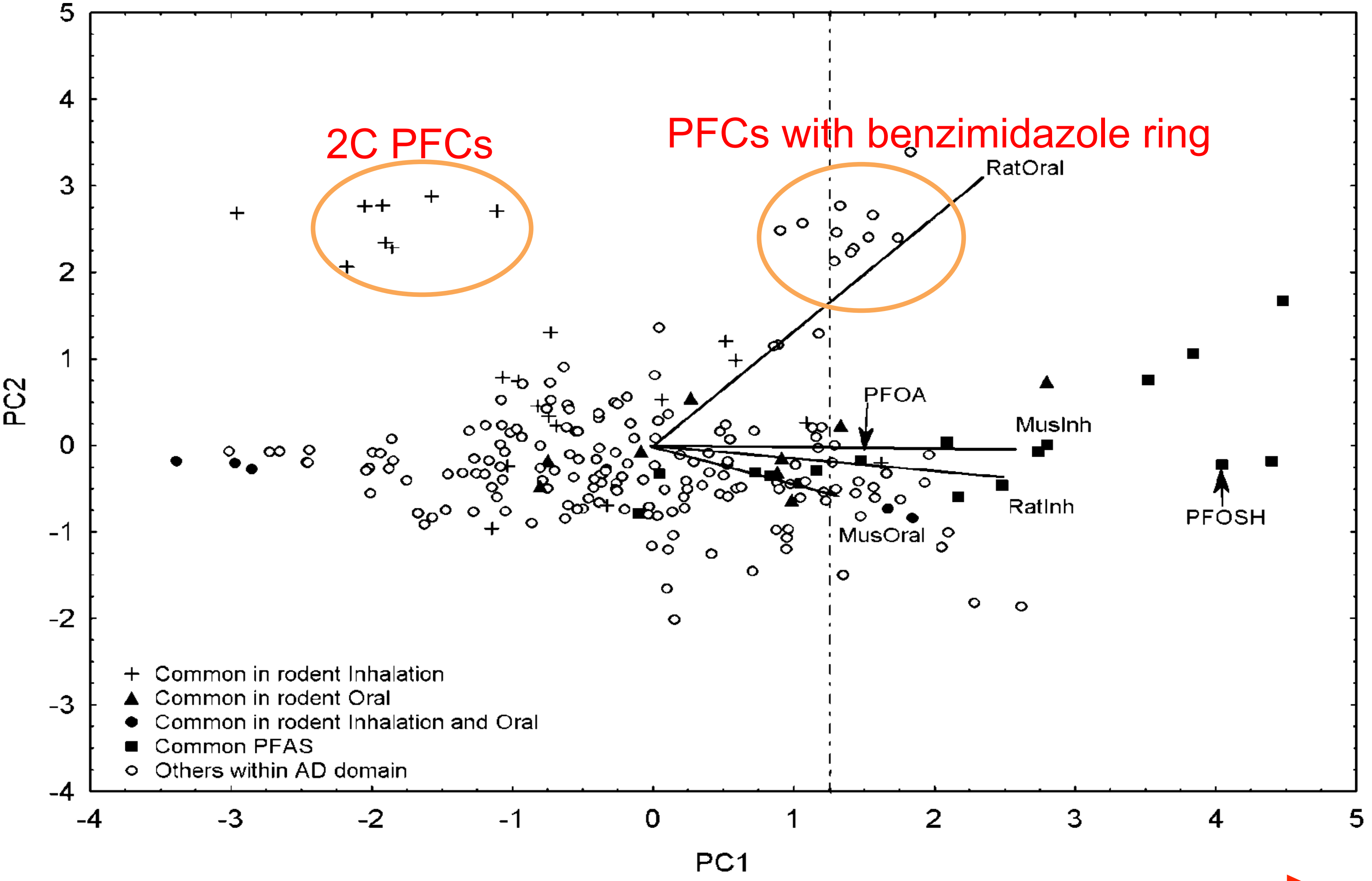
AD study on 250 PFCs in the REACH Pre-Reg. List





- **75.6% coverage of Mouse model: 61 compounds are out of domain**
- **78.8% coverage of Rat model: 53 PFCs out of AD.**

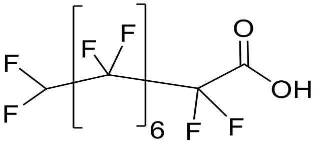
PCA plot for cumulative toxicity trend



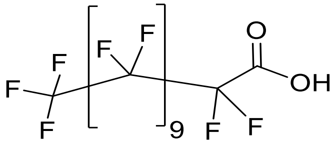
Increasing Cumulative Toxicity (PC1 EV%: 75.6%)

Dr. Ester Papa
QSAR Research Unit - DiSTA - University of Insubria - Varese (Italy)

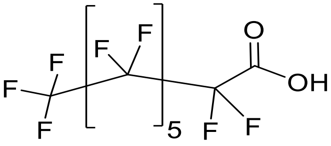
PFCs prioritization based on cumulative toxicity



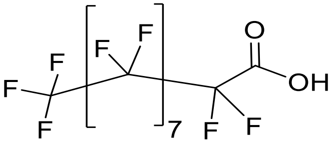
76-21-1



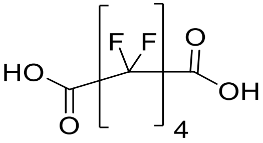
307-55-1



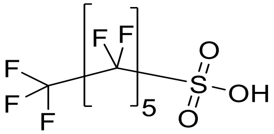
335-67-1



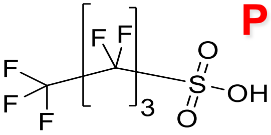
335-76-2



336-08-3

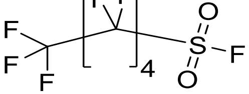


355-46-4

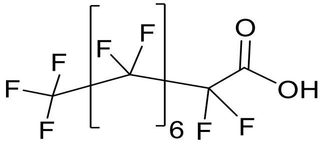


375-73-5

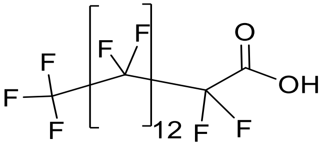
PFOA



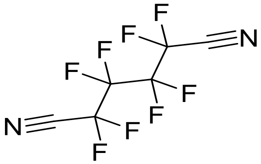
375-81-5



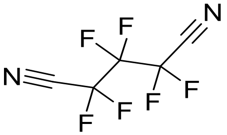
375-95-1



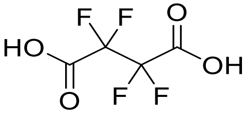
376-06-7



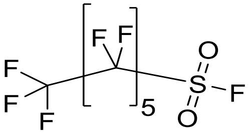
376-53-4



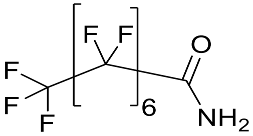
376-89-6



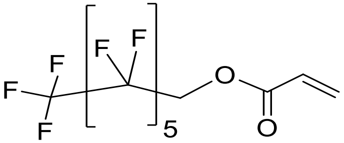
377-38-8



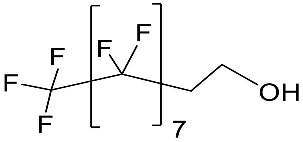
423-50-7



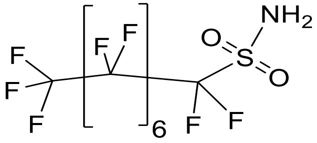
423-54-1



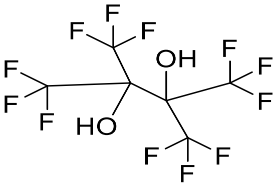
559-11-5



678-39-7

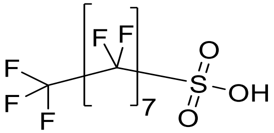


754-91-6



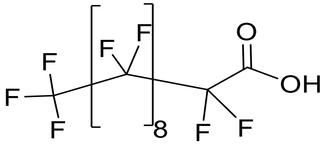
918-21-8

These chemicals were suggested to the CADASTER Partners for experimental tests

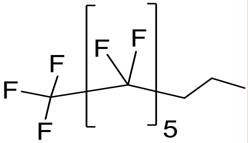


1763-23-1

PFOSH



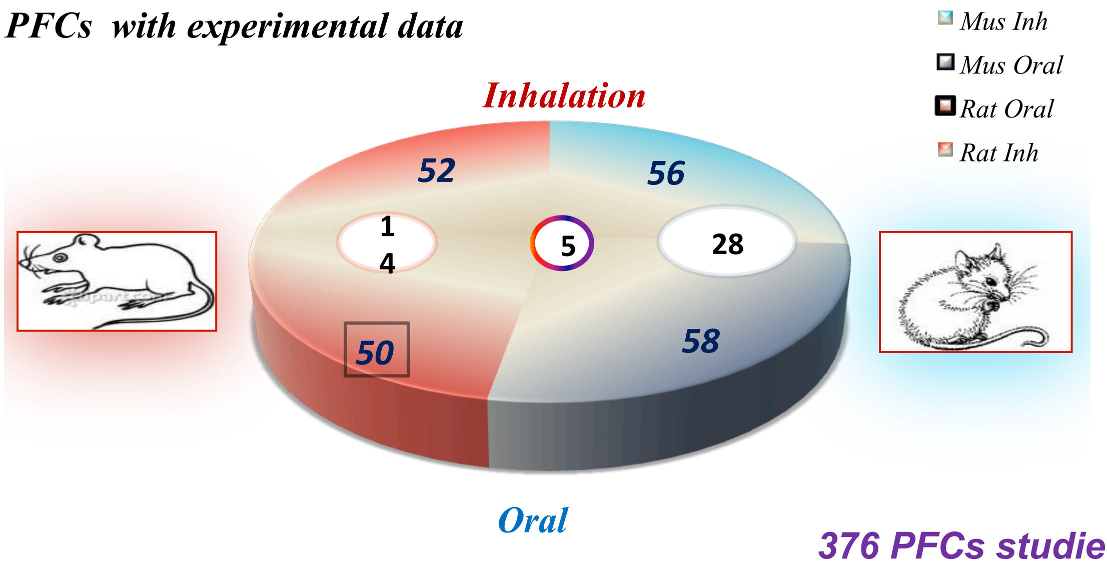
2058-94-8



17527-29-6

Dr. Ester Papa
QSAR Research Unit - DiSTA - University of Insubria - Varese (Italy)

PFCs toxicity in Rodents: integration of results



376 PFCs studied

- Starting from 50-58 experimental data, individual, externally predictive, models were applied for predictions of 250-376 PFCs in ECHA list for REACH (structural AD coverage of QSAR models: 75.6-90.9%)
- 22 PFCs prioritized by cumulative toxicity trend (PCA)

Take home message

1. Follow the OECD Principles

2. Have clear why you are building/applying your QSAR model → Descriptive – Predictive QSAR

3. QSAR Is not a “competition” → Consensus approach

Thanks To: QSAR Research Group - Insubria 2009-2012

Paola Gramatica, Ester Papa,
Simona Kovarich,Barun Bhhatarai,
Jiazhong Li, Mara Luini, Nicola Chirico,
Stefano Cassani, Elisa D'Onofrio,
Partha Pratim Roy,
Lidia Ceriani, Leon van der Wal

Physico-chemical property prediction of emerging pollutants: PFC and (B)TAZ for environmental distribution
QSAR and QSPR models for emerging pollutants: WP3 activities within the FP7 European Project CADASTER

Thank you for your attention

http://www.qsar.it