QSAR studies on mouse inhalation LC_{50} data of perfluoro-compounds

Barun Bhhatarai, Ester Papa, Paola Gramatica

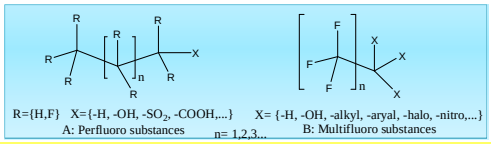
QSAR Research Unit in Environmental Chemistry and Ecotoxicology, DBSF, University of Insubria, Varese, Italy.

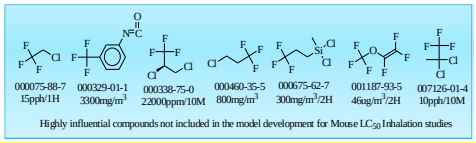
e-mail: barun.bhhatarai@uninsubria.it

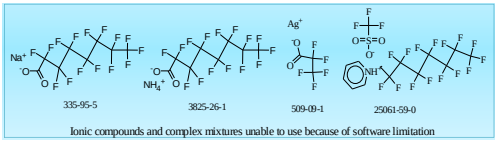
INTRODUCTION

Perfluorinated compounds have been very useful for chemical and mechanical industries mainly due to the stability of C-F bond, hydrophobic effect and thermal resistivity, but the fluorinated residues released in the environment are not quite degradable and are leading to environmental bioaccumulation, bio-persistence and health problems [1]. The mammalian toxicity of some of Perfluoroalkylated (PFA) compounds are studied in laboratory [2] but there are very few data available for developing Quantitative Structure-Activity Relationship (QSAR). Analysis of the structure of these compounds indicate that many congeneric fluorinated analogs are available and can be present in the environment. There is a general agreement that similarity in structure and properties indicates similar environmental fates and human health concerns. Thus, a dataset was prepared to study several per- and multi- fluorinated compounds with different functional groups as alcohols, acids and esters.

QSAR modelling by Multiple Linear Regression (MLR) is used as an *in silico* screening tool to predict the Mouse Inhalation (LC_{50}) toxicity data of a combined set of 56 per- and multi- fluorinated compounds [3]. The best model obtained for two splitting criteria and the Full model applicable to total 250 perfluoro compounds for which LC_{50} inhalation data were not found are reported. The models were developed according to the OECD principles. The performance of the QSAR models, the descriptors involved, the mechanistic interpretation are discussed. The applicability domain of the model and the predicted (Mouse Inhalation LC_{50}) response value based on the best selected models will be presented.







MATERIALS and METHODS

DATA SET The 56 experimental data were taken from the ChemIDplus Advance database pertaining to Mouse Inhalation (LC_{50}) toxicity. The set includes per- and multi- fluorinated compounds that are widely used and found in environmental matrices.

Additional per-fluorinated compounds whose toxicity data was not available were added resulting in a dataset of 250 compounds in total

VERY FEW DATA

QSAR model based on 56 compounds is used to define the 'applicability domain' on entire set of 250 compounds along with their prediction for Mouse LC_{50} inhalation.

MOLECULAR DESCRIPTORS

537 molecular descriptors (0D; 1D; 2D) were calculated by the DRAGON [3] from the XYZ coordinates in Hyperchem [4] minimization (AM1)

MULTIPLE LINEAR REGRESSION MODELS on descriptors selected by Genetic-Algorithm based Variable Selection method was performed by *Ordinary Least Squares* regression (OLS) [5].

EXTERNAL VALIDATION

Prediction set selection based on splitting by a) the structure similarity (by Kohonen Maps - Artificial Neural Networks (K-ANN)) & b) the Random Selection through activity sampling [6].

TOOLS OF VALIDATION AND DIAGNOSTICS

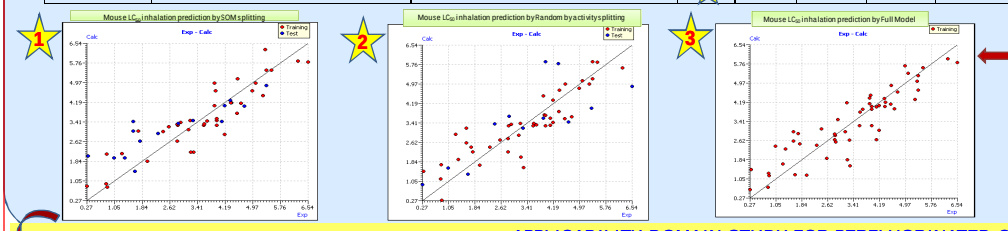
Models were developed following the 'OECD principles for QSAR validation' [7].

- Internal (by Q^2_{LOO} and Q^2_{BOOT}, Y-scrambling) and external validation (verified by Q^2ext).
- The quality of the best models were checked by Residuals and Williams plot
- Applicability Domain (AD for 250 PFCs) was verified by leverage approach.
- Mechanistic explanation based on the selected descriptors

RESULTS AND DISCUSSIONS

Response (Compounds)	Splitting	Compounds	Variables selected	Model	R^2	Q^2_{LOO}	Q^2_{BOOT}	Q^2 ext
Mouse Inhalation LC_{50} (56)	Self Organizing Map 28.5%	Train: 40 Test: 16	TPSA(NO); H-048; nCrs; MATS1v	1	*80.86*	*75.41*	*71.93*	*72.07*
	Random by Activity 20%	Train: 44 Test: 12	H-048; nHDon; F-082; X1sol	2	*77.17*	*72.08*	*70.09*	*61.00*
	Full model		MLOGP; X3v; F01[C-C]; H-048	3	*79.83*	*76.31*	*75.38*	-

Splitting	R^2	Q^2_{LOO}	Q^2_{BOOT}	Q^2ext
Full model	*78.67*	*74.68*	*72.75*	-
Random by Activity	*74.56*	*69.68*	*66.82*	*87.50*

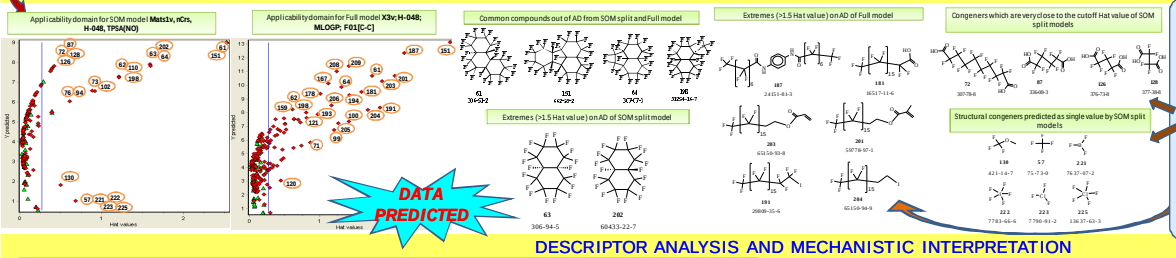


$$log\ 1/LC_{50} = 4.21 - 1.27\ (\pm 0.31)\ MlogP + 1.43\ (\pm 0.46)\ X3v + 0.38\ (\pm 0.13)\ F01[C\text{-}C] - 1.14\ (\pm 0.37)\ H\text{-}048$$

n=56, s=0.717, K_x=42.34, K_{xy}=50.40, $RMSE_{train}$=0.68, $R^2(Y_{scram})$=0.394

- Self Organizing Map (SOM) split model was more robust and predictive than Random by activity split model
- Full model descriptors are as predictive and robust as SOM split model
- Set of descriptors obtained by SOM split model and Full model performed better in different split sets

APPLICABILITY DOMAIN STUDY FOR PERFLUORINATED COMPOUNDS



- AD of SOM split model covers more compounds than that of Full model
- common compounds that are out of AD are polycyclic Perfluoro Alkylated compounds
- di-acids are predicted close to the domain by SOM model
- SOM predicted structural analogs of inorganic skeleton with same value
- Extremes compounds (>1.5 HAT value) of Full model are more than 15 C-atom long and probably extrapolated as the longest compounds in our model are of 7 carbon atoms

DESCRIPTOR ANALYSIS AND MECHANISTIC INTERPRETATION

Full model without splitting

- MLogP (–0.81) = Moriguchi octanol-water partition coefficient = hydrophobicity
- X3v (0.60) = presence of heteroatom and double/triple bonds present in the compound
- F01[C-C] (0.48)= Frequency of C-C at topological distance 01
- H-048 (–0.42) = H attached to C2(sp^3)/C1(sp^2)/C0(sp) carbon

Self Organizing Map (SOM) based splitting

- TPSA(NO) (0.59) = Topological polar surface are a using N, O polar contribution
- H-048 (–0.47) = H attached to C2(sp^3)/C1(sp^2)/C0(sp) carbon
- nCrs (0.33) = number of ring secondary C(sp^3) [only for 3 compounds]
- MATS1v (0.20) = Moran autocorrelation – lag 1/ weighted by atomic van der Waals volumes

Mechanistic interpretation

- nCrs = present to fit cyclohex-1-ene and cyclopent-1-ene ring containing compounds (CAS# 336-19-6, 706-79-6 & 775-40-6), when excluded these 3 compounds and redeveloped the SOM model, the third best model was with the same set of descriptors as Full model. nCrs helped in enlarging the AD encorporating unsaturated ring compounds
- H-048 = the formal oxidation number of hydrogenated carbon attached to the electronegative atoms, its importance in both models suggests its role in prediction as well as fitting
- X3v and MATS1v (r^2=0.514) representing the electronic property and the volume of the compounds involved
- X1sol (0.30) = solvation connectivity index chi 1 = representing enthalpies of non-specific solvation
- F01[C-C] = corresponds to chain length, as the alkyl chain length increases the C-C increases giving higher values to the longer chain

Increasing chain length (F01[C-C]), increase in the bond order as well as presence of heteroatom (X3v) has positive contribution increasing LC_{50} inhalation toxicity
Increasing the hydrophobicity (MlogP) and increase in oxidation of carbon (H-048) to higher value has negative contribution decreasing LC_{50} inhalation toxicity

Future work 1) comparison with Rat Inhalation data for finding similarities and differences understanding Inter species toxicity studies of Perfluoro compounds
2) QSA(P)R model also including ionic (Na^+, K^+, NH_4^+, Li^+) perfluoro compounds and modeling of toxicity data belonging to them

REFERENCES

1) Hekster F.M. et al. *Rev. Environ. Contam. Toxicol.* **2003**, 179, 99-121. 2) Renner R. *Environ. Sci. Technol.* **2003**, *37*, 201A. 3) Todeschini R., Consonni V. and Pavan M., 2001. *DRAGON – Software for the calculation of molecular descriptors*, rel. 1.12 for Windows. Free download available at http://www.disat.unimib/chm; 4) HYPERCHEM. Release 7.03 for Windows, 2002. Hypercube, Inc. Florida, USA. 5) Todeschini, R., 2001. *Moby Digs - Software for multilinear regression analysis and variable subset selection by Genetic Algorithm*, rel. 2.3 for Windows, Talete srl, Milan (Italy). 6) Gramatica, P. QSAR Comb. Sci. 2007, 26, 694-701. 7) OECD Guidelines, http://www.oecd.org/dataoecd/33/37/37849783.pdf.

(Financial support by European Union through the project CADASTER FP7-ENV-2007-1-212668)