Exploring the QSARs for OH Tropospheric Degradation of VOCs using freely available online descriptors

Partha Pratim Roy, Simona Kovarich, Ester Papa, Paola Gramatica

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, Varese (Italy)
E-mail: partha.roy@uninsubria.it, paola.gramatica@uninsubria.it

Introduction

Reactions with hydroxyl radical (-OH) is the most important pathway of day time removal of organic pollutant in atmosphere.

Three central OECD principles (2, 4 and 5) for QSAR (OH degradation) model validation is highlighted in this study with special emphasis specially model reproducibility (principle 2) by the users because of up gradation of commercial software with somewhat changed descriptors value or no more in the newer version

Additionally the models were applied to two set of CADASTER chemicals for their applicability domain and their predictions were compared with the widely used EPI Suite predictions

Materials & methods

Dataset: The experimental data of the OH radical degradation rate constants of 460 heterogeneous organic compounds were obtained from literature [1].

Descriptors: Zero, mono-, bi-dimensional descriptors available in DRAGON 5.5 [2]
2D descriptors available at CADASTER web[3]
ETA descriptors[4]
Quantum-chemical descriptors (HOMO, LUMO, Δ(HOMO-LUMO)) calculated by *HYPERCHEM* (AM1 method) were always added.

Method: Genetic Algorithm-Variable Subset Selection (GA-VSS), Multiple linear regression (MLR) and Ordinary Least Squares regression (OLS)

Data splitting: Random by response, K-ANN and K-means clustering

Statistical parameters: Internal (R^2, Q^2_{LOO}, Q^2_{BOOT}) external (Q^2-F1[5], Q^2-F2[6], Q^2-F3[7], ^2_m [8] concordance correlation coefficient [9])

Results and Discussion

Unambiguous Algorithm

The application of any QSAR models to chemicals without experimental data largely depends on model reproducibility

Regular up gradation of software almost abolish the reproducibility

Old model DRAGON 5.0: HOMO, nX, CIC0, **nCaH** → Model reproducibility → ?

Absent by name/somewhat different values

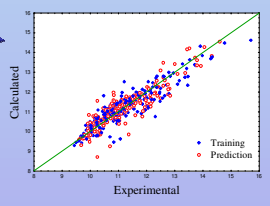
Serious drawbacks for QSAR modeling both for model developer and users and need for *QSAR model up gradation*

New models from updated and reproducible descriptors

DRAGON 5.5+MOPAC: HOMO, nX, IDE, **nCbH**

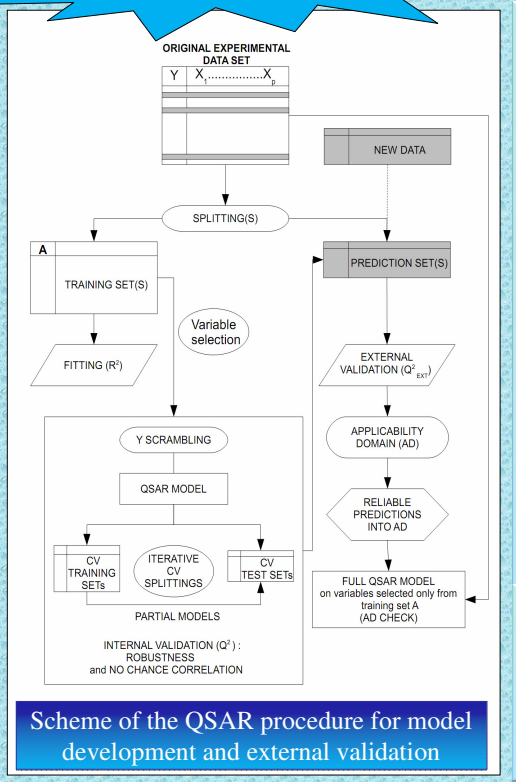
Identical or with almost similar information of those in the previously published model (HOMO, nX, CIC0, nCaH)

Similar mechanistic meaning and prediction



Online+MOPAC: HOMO, SeaC2C2aa, D_path(F, rel), G_(Cl, I, Br)

Confirm the ability of *GA* to select similar meaningful descriptors from different input as well as training set composition

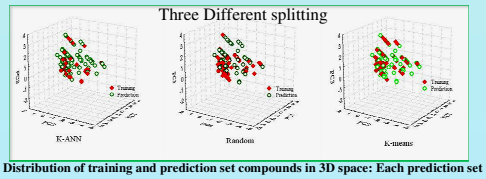


Scheme of the QSAR procedure for model development and external validation

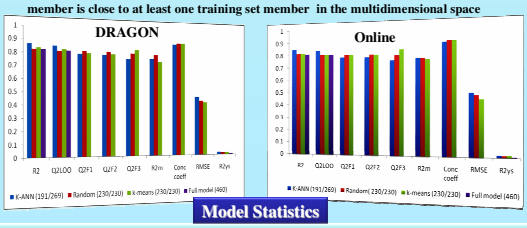
Validation

Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose.

QSAR → Statistically Robust *Internally & Externally* → *Accurate prediction* of chemicals *without Exp. activity* → *Validity & Reliablity* → O K



Distribution of training and prediction set compounds in 3D space: Each prediction set member is close to at least one training set member in the multidimensional space



Model Statistics

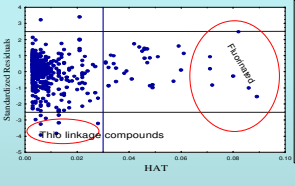
Models selected by GA from three different splitting procedures (K-ANN, K-means, random) demonstrated high performance in predicting external chemicals of different typology, having same set of descriptor combination

The developed splitted as well as the full models are statistically significant both internally as well as externally

Models obtained from both DRAGON and online descriptors have comparable statistics

Results of randomization (R^2_{ys}) indicate that the models are not by chance

Applicability Domain to CADASTER chemicals



AD was defined by the leverage approach, identification problematic chemicals (structural and response outliers)

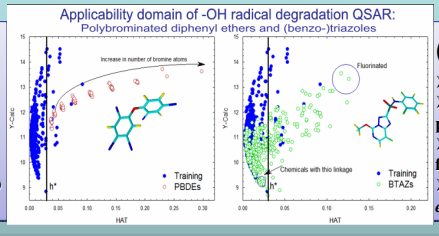
Rsponse

Overestimated: Triethyl phosphate, 2-(chloromethyl)-3-chloro-1-propene

Underestimated: Bromomethane , Dimethylsulfide, Diethyl sulfide, Ethyl methyl sulfide, 3-methyl-1,2 butadiene

Structural

Fluorinated compounds



PBDEs

- Outside AD
- ↑ No of Br tendency to go out of AD
- 94% correlation between Our and EPI Suite prediction
- Additional information on AD from our model

(B)TAZs

- 75% chemicals inside AD
- No correlation with EPI Suite prediction
- Compounds with thio linkage found to be toxic
- Information on AD regarding *extrapolation or interpolation*

Comparison of the present models with previously published QSARs

Reference	Techniques	No. Des.	N.Comp (T/P)	Descriptors	Q^2_{LOO}	RMSE Tr	Q^2_{Ext}	RMSE Ext
Bakken &Jurs	CNN	5	52/5	topological		0.07		0.06
	CNN	10	281/31	topological, electronic,		0.23		0.25
Gramatica et al.	MLR	4	234/226	HOMO, nX, nCaH, CICO	0.82	0.42	0.81	0.44
Gramatica et al.	MLR	6	460	HOMO, MATS1m, nDB, nO, CIC2, RTeþ	0.84	0.41		
Öberg et al	PLS	333/7	495/238	-	0.88	0.45	0.84	0.50
Wang et al.	MLR	4	44/11	HOMO, Q_H, MSA andμ	0.81	0.14	0.92	0.08
Wang et al.	PLS	22/3	576/146	(Ds, HOMO,nX, BELm2)	0.87	0.39	0.87	0.43
Present study	MLR	4	460	HOMO,nX, nCbH, IDE	0.82	0.43		
Present study	MLR	4	460	HOMO, SeaC2C2aa, D_path(F, rel), G_(Cl, I,Br)	0.80	0.45		

- Perfect comparison not possible because of use of different dataset and algorithm.
- Similar **mechanistic meaning** from different models

Conclusions

- Up gradation of useful OH model with similar mechanistic meaning
- Reproducibility of the model incresed by using online freely available descrptors
- Identification of problematic chemicals and the possible comment
- Prediction of CADASTER chemicals (PBDEs and (B)TAZs) with special emphasis on their applicability domain

References

1.Atkinson, R. J Phys Ref Data 1989, Monograph 1
2.DRAGON for Windows, ver.5.5, Talete srl, Milano, Italy.
3.www.cadaster.eu/databases
4.Roy, K.; Ghosh, G. Internet Electron J Mol Des 2003, 2, 599-620
5.Shi *et al.* J Chem Inf Comput Sci 2001, 41, 186-195.
6.Schüürmann *et al.* J Chem Inf Model 2008, 48, 2140-2145.
7.Consonni, V.; Ballabio, D.; Todeschini, R. J Chem Inf Model 2009,49, 1669-1678.
8. Roy,P.P.; Roy, K. QSAR Comb Sci 2008, 27, 302-313.
9. Chirico, N.; Papa, E.; Gramatica, P. Presented at the 21 SETAC Europe Meeting, May 2011, Milan, Italy.

The work is in press in ***Journal of Computational Chemistry***

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