QSPR Models for Predictions and Data Quality Assurances: Melting Point and Boiling Point of Perfluorinated Chemicals

www.cadaster.eu

Barun Bhhatarai[#,*], Wolfram Teetz[δ], Tomas Öberg[†], Tao Liu[†], Nina Jeliazkova[‡], Nikolay Kochev[§], Ognyan Pukalov[§], Igor V. Tetko[δ], Simona Kovarich[#] and Paola Gramatica[#]

[#]QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology (DBSF), University of Insubria, via JH Dunant 3, Varese, 21100, Italy. Email: paola.gramatica@uninsubria.it; [†]School of Pure and Applied Natural Sciences, Linnaeus University, SE-391 82, Kalmar, Sweden. Email: Tomas.Oberg@lnu.se; [‡]Ideaconsult Ltd, 4 A. Kanchev str., Sofia 1000, Bulgaria; Department of Analytical and Computer Chemistry, University of Plovdiv, 24 Tsar Assen Str., Plovdiv 4000, Bulgaria. Email: jeliazkova.nina@gmail.com; [δ]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen - German Research Center for Environmental Health, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany. Email: itetko@vcclab.org.

ABSTRACT

Quantitative structure-property relationship (QSPR) studies on Melting Point and Boiling Point of Perfluorinated Chemicals (PFCs) are presented. PFCs are studied under the EU-FP7 funded CADASTER project to understand its behavior in biota and environment. They are considered as 'emerging pollutants' and found widely distributed in the environment, released due to their widespread use in different household and industrial products as cleansers, fire-fighting foams, micelles, repellants for leather, paper, and textiles etc. Continues exposure of these chemicals is found to be the source of bio-accumulation in body parts of human, wildlife and is ultimately becoming the cause of toxic reactions and poisoning.

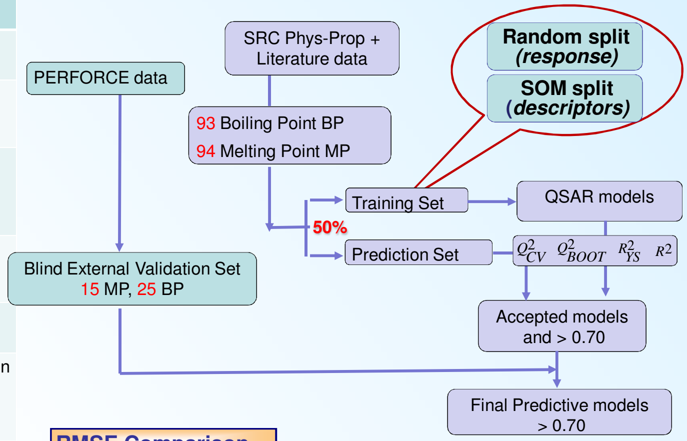
Models are developed using SRC PhysProp data as described below. In addition, the predictive performances of the developed models were verified on a blind external validation set (EV-set) prepared from experimental values available from PERFORCE database. This database contains only long chain perfluoro-alkylated chemicals, particularly monitored by regulatory agencies like US-EPA and EU-REACH. QSPR modeling using different approaches, internal and external validation on two different prediction sets and studies of the applicability domain highlight the robustness and high accuracy of the proposed models. Finally, Melting Point for additional 397 PFCs and Boiling Point for 364 PFCs for which experimental measurements are unknown were predicted, verifying their applicability domain. The set of descriptors which best describes the structure-property relationship, the similarities, and the differences observed will be discussed as well as the consensus model predictions.

MATERIALS AND METHODS

	HMGU, Germany	IDEA Consult, Bulgaria	UI , Italy	LNU , Sweden
Descriptors	E-State indices [1]	Fragement based	DRAGON (0D – 2D) [2]	
Descriptor Selections	Pearson Pairwise Correlation	Exhaustive isomorphism search of fragment against structure	Pearson Pairwise Correlation & Genetic Algorithm	variable influence on projection (VIP)

Descriptors used for Modeling	MP = 87 indices
	BP = 66 indices
		MP = 3 descriptors
		BP = 8 descriptors
			MP = 4 descriptors
			BP = 4 descriptors
				MP = 37 descriptors at 3 components
				BP = 149 descriptors at 4 components

Methods	Associative Neural Network (ASNN) [3] Architecture: 10x3x1	Multiple Linear Regression (MLR) using ordinary-least-squares (OLS)		Partial least squares regression (PLSR)
External validation	Double: Prediction sets by splitting and blind External Validation set			Single: External Validation Set
Structural Applicability Domain	Distance to model (DM) on standard deviation of ensemble prediction, 5xcross-validation	Williams plot for response outliers Leverage approach (H matrix) for structural chemical domain [4, 5]		residual standard deviation (Euclidean distance) and leverage (Mahalanobis distance [6])



PERFORCE data

SRC Phys-Prop + Literature data

93 Boiling Point BP

94 Melting Point MP

Random split *(response)*

SOM split *(descriptors)*

50%

Training Set

Prediction Set

QSAR models

Q^2_{CV} Q^2_{BOOT} R^2_{YS} $R2$

Blind External Validation Set 15 MP, 25 BP

Accepted models and > 0.70

Final Predictive models > 0.70

Data Quality Assurance

CAS	Endpoint reported	Data from PhysProp (°C) used by UI, LNU, IDEA	UI Predictions	LNU Predictions	IDEA Predictions	Data (°C) used by HMGU [7]	HMGU Predictions
76-16-4	MP	–101.00	–155.01	–138.33	-154.67	–155.60	-111.66
307-34-6	MP	–42.0	–29.65	–54.73	-43.58	–56.80	-57.45
354-32-5	MP	146	-8.11	-91.56	-40.85	–146.0	-86
423-55-2	MP	<25	-4.11	-40.99	-27.71	–6.0	-59.17
1493-13-6	MP	<25	37.76	–31.38	14.82	–40.0	-12.57
426-65-3	MP → BP	75.5	53.87	-21.43	-0.003	n/a	n/a
355-46-4	BP	238.5	227.69	241.87	212.34	225.0	217.02
375-73-5	BP	211.0	195.62	207.33	182.77	200.0	191.36

RMSE Comparison

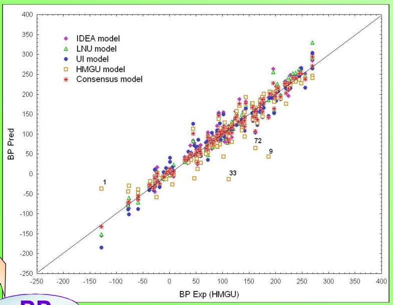
	RMSE (training set)				
	EPI	UI	LNU*	HMGU	IDEA
Melting Point (94)	47.97	42.42	31.12	36.48	40.67
Boiling Point (93)	24.80	21.39	14.12	31.89	17.25

	RMSE (PERFORCE)				
	EPI	UI	LNU*	HMGU	IDEA
Melting Point (15)	n/a	27.19	26.54	34	38
Boiling Point (25)	n/a	30.32	21.92	22	23

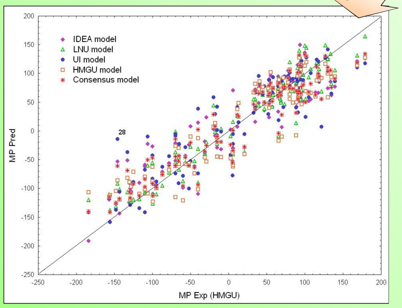
*LNU model was developed without external validation

RESULTS AND DISCUSSION

BP



MP



MP Model	UI			LNU		HMGU		IDEA		
	R^2	Q^2_{LOO}	range Q^2_{EXT}	R^2	Q^2_{LOO}	R^2	R^2_{EXT}	R^2	Q^2_{LOO}	range Q^2_{EXT}
SOM split	0.83	0.79	0.61-0.76	0.90	0.85	0.80	0.75	0.86	--	0.61-0.76
Random split	0.84	0.80	0.73-0.76	0.90	0.84	0.81	0.75	0.84	--	0.72-0.76
FULL model	0.80	0.78	--	0.89	0.86	0.85	--	0.80	0.78	--

BP Model	UI			LNU		HMGU		IDEA		
	R^2	Q^2_{LOO}	range Q^2_{EXT}	R^2	Q^2_{LOO}	R^2	R^2_{EXT}	R^2	Q^2_{LOO}	range Q^2_{EXT}
SOM split	0.95	0.94	0.86-0.91	0.98	0.94	0.79	0.58	0.95	--	0.92-0.95
Random split	0.94	0.92	0.93-0.94	0.98	0.93	0.78	0.57	0.96	--	0.93-0.94
FULL model	0.94	0.93	--	0.97	0.94	0.85	--	0.95	0.94	--

CONCLUSIONS

- **Combination of different modeling approach also helps to replenish the inability of one model with the support of another.**
- **The results fit our experience that a consensus model, built from independently developed models using different descriptors and using different algorithms, delivers the best prediction results.**
- **In the special case of PFCs, simple statistical algorithms applied to complex descriptors perform about as good as complex algorithms applied to simple descriptors. Developing both types of models enables a more specialized and also more detailed look on outliers and opens lots of possibilities to analyze them.**
- **Chemical interpretation of and experimental design emerging from the models benefit from having a set of models representing different views of the underlying mechanics.**
- **The data collected from the database has a high number of errors like mixed up algebraic signs or approximated values, so that data validation and overlap is necessary. Our approach which deals with the relation between BP and MP gives valuable information that can be employed and is also robust against erroneous data.**

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

REFERENCES

[1]ESTATE: Kier, L. B.; Hall, L. H. An Electrotopological State Index for Atoms in Molecules. Pharm. Res. 1990, 7, 801-807.

[2]DRAGON: TALETE srl, Via V. Pisani, 13 - 20124 Milano - Italy

[3]ASNN, Tetko, I. V. Associative neural network, Neural Processing Letters, 2002, 16, 187-199.

[5]MAHALANOBIS, Mahalanobis, P C (1936). "On the generalised distance in statistics". Proceedings of the National Institute of Sciences of India 2 (1): 49–55. http://ir.isical.ac.in/dspace/handle/1/1268.

[6]AD, Jaworska, J., Nikolova-Jeliazkova, N., Aldenberg, T. (2005). QSAR applicability domain estimation by projection of the training set in descriptor space: A review. ATLA-Alternatives to Laboratory Animals 33(5): 445-459.

[7]AD, Papa, E.; Kovarich, S.; Gramatica P. Development, Validation and Inspection of the Applicability Domain of QSPR Models for physico-chemical properties of Polybrominated DiphenylEthers QSAR and Combinatorial Science, 2009, 28, 790-796.

[7] http://www.chemicalbook.com/; http://www.sigmaaldrich.com/