



Boiling points of halogenated aliphatic compounds:

A quantitative structure-property relationship for prediction and validation

Tomas Öberg

Kalmar Univ., Dept. Biol. Environ. Sci. Sweden.

Email: tomas.oberg@hik.se or info@tomasoberg.com

Presentation at the Third Indo-US Workshop on Mathematical Chemistry, August 2 - 7, 2003, Duluth, Minnesota, USA



The boiling point

- An important physicochemical property
 - Identifies compounds, assess purity
 - Assesses volatility
 - Estimates other properties, e.g. vapor pressure
 - An indication of attractive forces between molecules
 - These intermolecular forces are directly related to the structure, and hence the boiling point may be correlated with structure
-

Halogenated aliphatics

- Used as refrigerants, blowing agents, anesthetics, solvents, fire extinguishers, etc.
 - The high strength of the fluorine-carbon and chlorine-carbon bonds make most of these compounds extremely stable
 - Some have a significant ozone depletion (ODP) and/or greenhouse warming potential (GWP)
 - The Montreal protocol currently controls 96 halogenated aliphatics (C₁-C₃) that are damaging to the ozone layer
 - Hydrofluorocarbons and perfluorocarbons are controlled by the Kyoto protocol to the Convention on climate change
-

Lack of data

- The boiling points have been determined for many halogenated aliphatics, but there are also gaps in the available data
 - Several reports have appeared regarding the estimation of the boiling point for short-chained halogenated aliphatic compounds
 - The development of estimation methods is however always limited by the availability of reliable calibration data
 - A few years ago Imperial Chemical Industries (ICI) released an additional set of data, which is used in this investigation
-

Scope

- Model the experimentally determined normal boiling points (NBP), i.e. at a pressure of 101.3 kPa, of halogenated aliphatic compounds from computationally derived molecular descriptors
 - Compare this approach with a group contribution method
 - Investigate if the developed QSPR model can be used to validate the available experimental database
-

Calibration data

- Experimentally determined values for the normal boiling points of 240 halogenated aliphatic compounds were obtained from the literature
 - A. L. Horvath (2001): Boiling points of halogenated organic compounds. *Chemosphere* **44**, 897-905
 - Experimentally determined boiling points were also obtained for 74 out of these 240 compounds from another data source
 - The PhysProp Database (Syracuse Research Corporation, Syracuse, NY, USA).
-



Descriptor generation and estimation software

- The chemical structures were sketched and modeled using HyperChem v7.01 (HyperCube, Inc., Gainesville, Florida, USA)
 - Using the force-field routine MM+
 - 1175 empirical descriptors were generated for each structure using Dragon v3.0 (Milano Chemometrics and QSAR Research Group, University of Milano-Bicocca, Milano, Italy)
 - Boiling points were estimated with the group contribution method of Stein and Brown using MPBPWIN v1.4 (U.S. Environmental Protection Agency, Washington, DC, USA)
-



Other software

- The data analysis and multivariate calibrations were carried out with the software:
 - Matlab v6.5 (Mathworks Inc., Natick, Mass., United States)
 - Statistica v6.1 (StatSoft Inc., Tulsa, Okl., USA)
 - Unscrambler v7.6 SR-1 (CAMO ASA, Oslo, Norway)
-

Data analysis

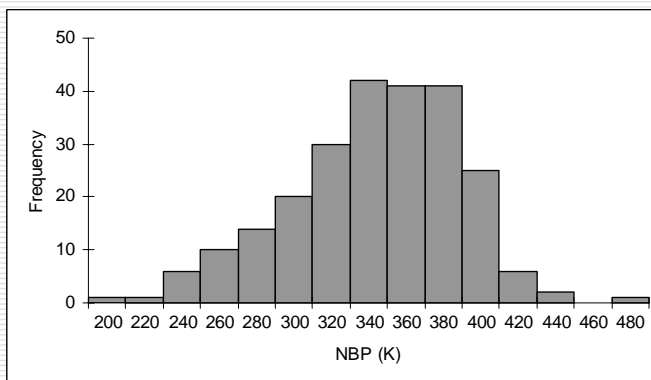
- Methods for data analysis and modeling:
 - Analysis of variance (ANOVA)
 - Principal component analysis (PCA)
 - Partial least squares regression (PLSR)
 - Implicit non-linear latent variable regression (INLR)
- Preprocessing
 - Auto scaling to zero mean and unit variance
 - Descriptor variables with minor influence in the PLS regression were assigned zero weight (identified with a jack-knife method)
- Validation
 - Cross-validation to establish the model rank
 - External test set to estimate the prediction error

Elemental compositions

Number of compounds with the different elemental compositions

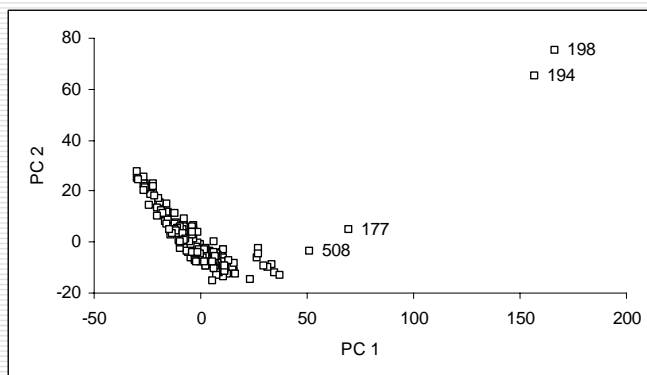
Atom type / No. atoms	0	1	2	3	4	5	6	7	8	9	>9
Fluorine	15	17	70	52	37	30	6	5	4	1	3
Chlorine	79	75	55	21	10	0	0	0	0	0	0
Bromine	148	65	23	4	0	0	0	0	0	0	0
Carbon	0	15	71	128	22	2	0	2	0	0	0
Hydrogen	33	59	49	42	31	13	7	3	3	0	0

Frequency distribution



Boiling points
in the range
191-462 K

Outliers from PCA and PLSR



Pentadeca-
fluoroheptane
(194), hexa-
decafluoro-
heptane (198)
are different.
17 other
outliers were
identified in a
preliminary
PLS regression

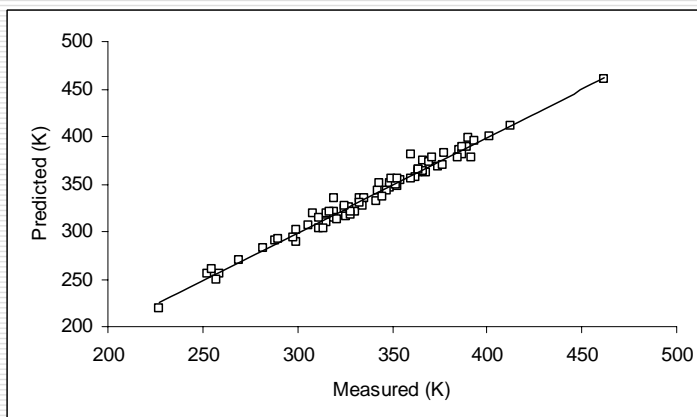
The calibration model

- The 221 objects remaining objects were randomly assigned to
 - a calibration set of 146 objects
 - a test set of 75 objects
 - 511 descriptor variables were selected for inclusion in the final model based on jack-knifing in preliminary runs
 - The number of latent variables to keep in the PLSR model was estimated to six using full (leave-one-out) cross-validation
 - The model was subsequently validated with the external test set
-

Model performance

- Calibration data (146 obj.):
 - SEC 4.90 K
 - R^2_{Cal} 0.989
 - External test set (75 obj.):
 - SEP 6.42 K
 - Q^2_{Ext} 0.976
 - Average absolute errors:
 - Calibration error 3.72 K
 - Test set prediction error is 4.78 K
 - Experimental error ~5-10 K
-

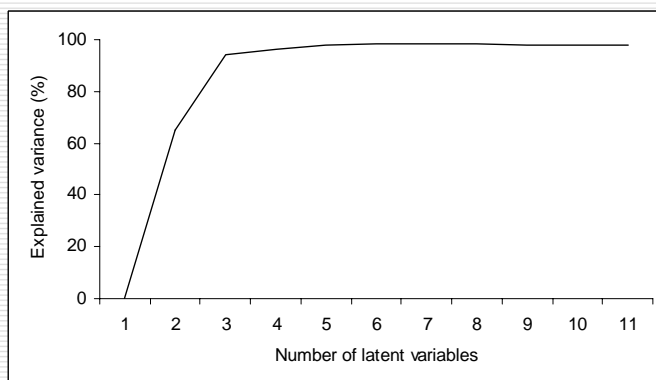
Predicted vs. measured (external test set)



Nonlinearities

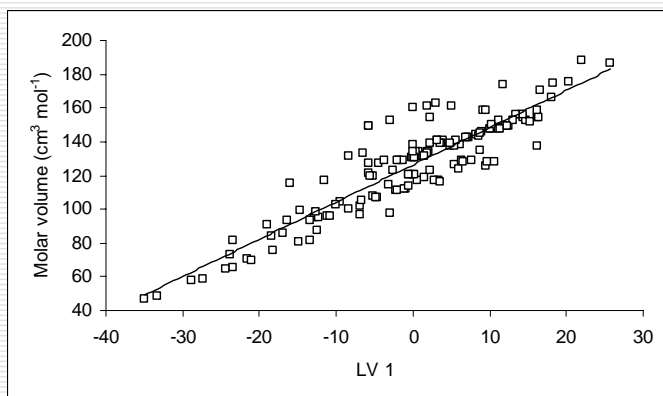
- Are nonlinear techniques like neural networks necessary to minimize the prediction error?
 - Recalibration with INLR, an extension of PLSR by expansion of the descriptor matrix with quadratic terms, did not improve the calibration model (SEC 5.02 K and R^2_{Cal} 0.989)
 - The already low prediction error also suggests that additional parameters or nonlinear approaches would not improve the outcome
- The bilinear PLSR models can in fact often accommodate nonlinear behavior by the addition of more latent variables

Two main factors



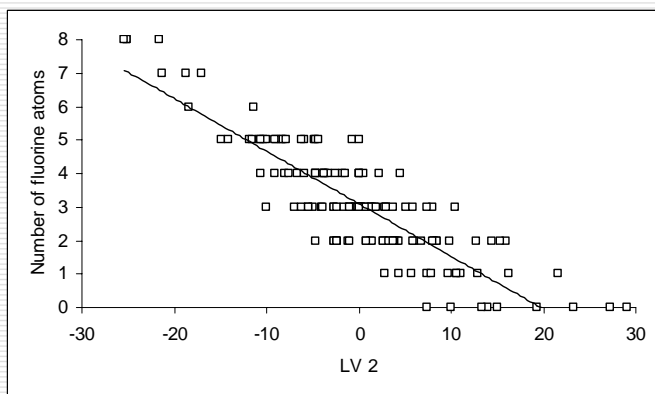
The first two latent variables describe most of the variation

The first latent variable



McGowan's molar volume (cm³ mol⁻¹) vs. the first latent variable

The second latent variable



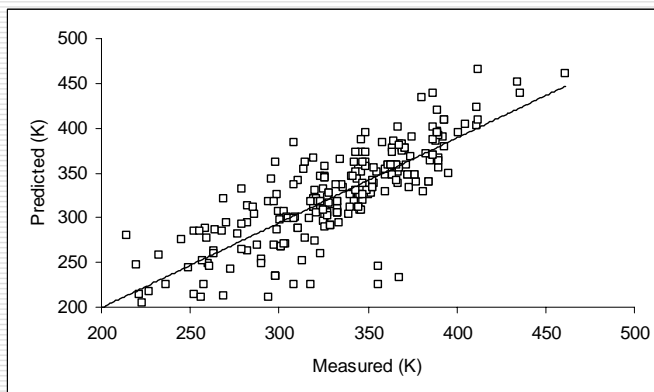
Fluorine atoms
(decreased
polarizability)
vs. the second
latent variable

ANOVA (explanatory model)

Boiling point described with linear model from size and decrease in fluorine substitution ($R^2=0.689$).

Source	SS (x100)	DF	MS (x100)	F	Prob.
Model	2222	2	1111	158.3	<0.00001
Error	1004	143	7.020		
Total	3226	145	22.25		
Factors					
Molar volume	2154	1	2154	306.9	<0.00001
No. Fluorine	764.9	1	764.9	109.0	<0.00001

The group contribution method (Stein & Brown)

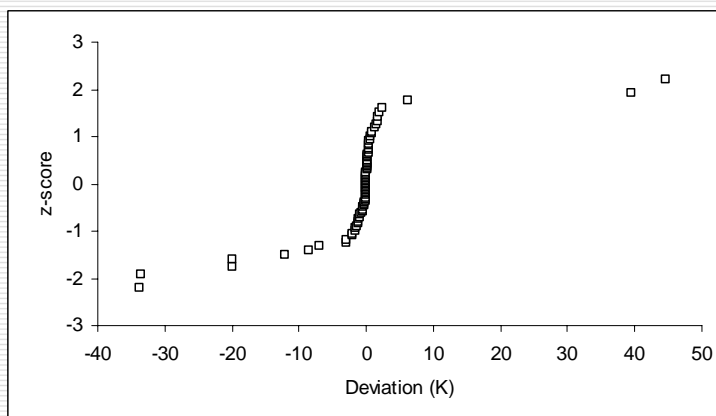


Predicted vs. measured normal boiling points (K) for all the 240 halogenated aliphatics (SEP 30.3 K and Q^2 0.664)

Validation of data

- Experimental data reported in the PhysProp Database for 74 compounds were used for comparison
- In general the agreement between data from the two sources was excellent, with a deviation between reported values of less than 5 K
- 7 objects did however deviate more than 10 K as seen from a normal probability plot

Deviation in NBP



The outliers again

- The frequency of deviations larger than 10 K between the two data sources is then 9%, so clearly there is a need for critical evaluation
- 19 outliers were initially identified in ICI study and it is of interest to examine these further
- Outliers can be due both to high variance or leverage in the descriptor variables or a high residual variance in the dependent variable
- Outliers in the descriptor variables, nos. 194 and 198, are outside of the calibration domain

Outliers – a comparison

No.	Compound	ICI (K)	PhysProp (K)	PLSR (K)	Group (K)
42	1,1,2-Trichloro-1,2-difluoroethane	296.2	-	346.3	342.2
62	1-Chloro-2,2-difluoroethane	326.2	308.3*	301.1	300.0
76	1,1,1,2,2,3,3-Heptafluoropropane	294.2	-	255.6	210.5
82	1,1,2,3-Tetrachloro-1,3,3-trifluoropropane	358.2	-	398.9	383.4
126	1,1-Dibromo-2,2-difluoropropane	348.6	-	388.6	394.8
154	2,3-Dichloro-hexafluoro-2-butene	381.2	341.7	347.4	328.2
177	Dodecafluoropentane	268.7	302.4	296.7	213.3
194	1,1,1,2,2,3,3,4,4,5,5,6,6,7,7-Pentadecafluoroheptane	368.2	369.2	385.6	233.4
198	Hexadecafluoroheptane	355.7	355.7	363.3	224.8

* Experimental value from Carlton.

Outliers – a comparison cont.

No.	Compound	ICI (K)	PhysProp (K)	PLSR (K)	Group (K)
308	1-Bromo-1,1-dichloro-2,2-difluoropropane	308.7	-	387.1	384.0
314	1,1,1-Trichloro-2,2-difluoropropane	326.2	346.2	364.2	343.7
363	1-Bromo-2-chloro-1,1,2,2-tetrafluoroethane	347.2	-	305.6	324.9
364	1,1,1-Tribromo-2,2,2-trifluoroethane	387.2	-	395.6	438.2
367	1,2,2-Trichloro-1,1-difluoroethane	389.7	345.1	345.7	355.8
393	1,1-Dichloro-1,2,2,3-tetrafluoropropane	346.2	-	390.6	386.3
408	1,1,1-Trichloro-2,2-difluoropropane	326.2	346.2	364.3	343.7
415	1,1,3-Tribromo-2,2-difluoropropane	412.2	-	458.7	465.3
421	1,2-Dichloro-2,3-difluoropropane	367.7	-	398.4	380.0
508	1,1,2,2,3,3,4,4,5-Nonafluoropentane	355.7	-	355.0	245.2

Outliers – a comparison cont.

- The results presented show that the PLSR model predictions agree well with the available experimental results from the PhysProp database and Carlton, with an average absolute error of 9.9 K
 - The deviation both in between the different data sources and with the group contribution method is much larger
 - It seems advisable to treat the reported experimental results for most of these outliers with caution
-

Validation of the ICI data

- The PLSR model predictions deviate more than twice the previously reported SEP for 21 compounds
 - 16 of these mismatching predictions belong to the outliers listed above
 - Three out of these 21 compounds have measured results reported in the PhysProp database that agree well with the ICI data, thus reducing the mismatch to 18 compounds
 - 92.5% of the ICI data were in good agreement, either with model predictions or with other experimental results
 - 96% of the PhysProp data can be validated using the same standards
-

Conclusions

- A bilinear PLSR model is a suitable choice to describe the quantitative structure-property relationship (QSPR) between the normal boiling points and computationally derived molecular descriptors for halogenated aliphatic compounds
 - The prediction error for a separate test set approaches the anticipated lower bound of experimental error
 - This is in agreement with the findings of other workers using similar approaches to study other groups of compounds
-

Conclusions cont.

- Using projection to latent variables instead of traditional methods of variable reduction, e.g. step-wise regression, facilitates the identification and removal of outliers
 - The physical-chemical interpretation of the main latent variables indicated as expected that the main source of variation can be attributed to molecular size and polarizability
 - The QSPR model developed in this study was more precise and accurate than the group contribution method of Stein and Brown
-

Conclusions cont.

- Within this domain the current QSPR model can be used to predict boiling points also for compounds where it is not available, or even for compounds not yet synthesized
 - QSPR models can be used to check and validate experimental data
 - The results presented in this study indicate that also here this is a viable approach
 - Comparison with experimental data from another source seems to support the accuracy of the model predictions
 - This modeling approach can readily be extended to other physical properties as soon as a sufficient amount of calibration data is available
-

Acknowledgements

- Support from the Faculty of Science, University of Kalmar, is gratefully acknowledged
-