Modeling the Property of Compounds from Structure: Statistical Methods for Models Validation

Sorana-Daniela Bolboacă[1] and Lorentz Jäntschi[2]

[1]„Iuliu Hatieganu“ University of Medicine and Pharmacy, 400023 Cluj-Napoca, Romania.
E-mail: sbolboaca@umfcluj.ro
[2]Technical University of Cluj-Napoca, 400020 Cluj-Napoca, Romania.

A molecular descriptors family on structure-property relationships study (MDF-SPR) was conducted in order to model the boiling points of alkanes using the compounds structure information.

The alkanes from C3 to C9 were included into study. Two MDF-SPR models, one with one descriptor and other with two descriptors, were identified. The estimation and prediction of the MDF-SPR models were analyzed. The methods used for validation of the obtained MDF-SPR models are presented.

The correlated correlation analysis was using in order to compare the performances of the obtained MDF-SPR models and of the MDF-SPR models comparing with previous reported model. The Steiger's Z test [1] at a significance level of 5% was applied.

The statistical analysis of the obtained MDF-SPR models demonstrated that the model with two descriptors has greater abilities in estimation and prediction compared with the model with one descriptor. More, the MDF-SPR model with two descriptors has greater abilities in estimation comapring with previous reported model. These observations were also sustained by the results of correlated-correlation analysis.

The multi-varied MDF-SPR model can be used in order to predict the property of interest of studied alkanes without any experiments and measurements, by using the MDF SPR Predictor application [2].

Acknowledgements

The research was partly supported by UEFISCSU Romania through projects ET36/2005 & ET108/2006.

References:

1. J. H. Steiger, Psychol. Bull. 87 (1980) 245.
2. ***, MDF SPR-SAR Predictor, © 2005, Virtual Library of Free Software [cited 2006 March]. Available from: URL: http://vl.academicdirect.org/molecular_topology/mdf_findings/sar

Ecomaterials and Processes: Characterization and Metrology
2007 April 19 – 21, St. Kirik, Plovdiv, Bulgaria

MODELING THE PROPERTY OF COMPOUNDS FROM STRUCTURE: STATISTICAL METHODS FOR MODELS VALIDATION

Sorana Daniela BOLBOACĂ[1] and Lorentz JÄNTSCHI[2]

[1] *„Iuliu Haţieganu“ University of Medicine and Pharmacy, 13 Emil Isac, 400023 Cluj-Napoca, Romania, http://sorana.academicdirect.ro*
[2] *Technical University of Cluj-Napoca, 15 Constantin Daicoviciu, 400020 Cluj-Napoca, Romania, http://lori.academicdirect.org*

ABSTRACT

A molecular descriptors family on structure-property relationships (MDF SPR) study was performed in order to model the boiling points of alkanes starting from the information obtained from the compounds structure. All alkanes from C3 to C9 was included in this study. Two MDF SPR models, one with one descriptor and other with two descriptors, were contsructed. The estimation and prediction of the models were analyzed in order to identify the best performing MDF SPR model. The best performing model was validated, and its correlation coefficient was compared with a previously reported model. The results of the study revealed that the MDF SPR approach is a useful method of model the boiling points of alkanes providing stable models.

INTRODUCTION

The boiling points of a sample of alkanes were previously studied by using 3D molecular descriptors. There were reported models with one, two, three, and four variables, respectively [1]. The equation of the best model (the model with three variables) had the following formula:

$$Bp(°C) = 727.26(\pm 20.76)\cdot 3D^0\chi - 19.46(\pm 0.9)\cdot 3DSRW2 + 7.99(\pm 0.39)\cdot M2 - 779.42(\pm 20.08) \quad (1)$$
$$n = 73;\ r = 0.9986;\ s = 2.17;\ F = 8340$$

where $3D^0\chi$ and 3DSRW2 are MIS (Method of Ideal Symmetry) indices, and M2 is a 3D modification of the Zagreb index; Bp(°C) is the boiling point.
A new method called molecular descriptor family on structure-property relationships (MDF-SPR) was been introduced by Jäntschi [2] proving its abilities in prediction of chemical compounds properties [3, 4].
The current study attempts to find the relationships between structure of alkanes with three, four, five, six, seven, eight and nine carbons and boiling points, to identify the best MDF SPR models and to analyze their estimation and prediction abilities.

DATA SET AND METHODOLOGY

A sample of seventy-three compounds, representing the isomers of the alkanes from three to nine (one C3 alkane, two C4 compounds, three C5 compounds, five C6 compounds, nine C7 compounds, eighteen C8 compounds and thirty-five C9 compounds, respectively) where included into the study.
The experimental data of boiling points were taken from a previously reported research [5]. The boiling points of alkanes have been investigated using MDF-SPR methodology [2]. The best performing mono and multivariate models were identified based on the value of the correlation coefficient. In order to demonstrate the absence of chance correlation into the obtained model, two validation methods were used (leave-one-out method, and leave-many out method).
Comparisons between correlation coefficients obtained by different models were analyzed by using the Steiger's Z test [6] at a significance level of 5%.

RESULTS

Two MDF SPR models with abilities in estimation of boiling points for studied alkanes were identify, one model with a single descriptor (see Eq.(2)) and other with two descriptors (see Eq.(3)), where $\hat{Y}_{1D}$ is the estimated boiling points by the model with one descriptor (Eq.(2)), $\hat{Y}_{2D}$ the estimated boiling points by the model with two descriptors, and *lbMdsHg*, *lGDrtGt*, and *lbDrfHt* are molecular descriptors.

Model with one descriptor

$$\hat{Y}_{1D} = -507.95 + 188.40\cdot lbMdsHg \quad (2)$$

Model with two descriptors

$$\hat{Y}_{2D} = -129.20 - 67.45\cdot lGDrtGt + 4.89\cdot lbDrfHt \quad (3)$$

Statistical characteristics in terms of squared correlation coefficients, Fisher parameters and associated significance, standard error of the MDF SPR models are presented in table 1. Analyzing the obtained MDF SPR models revealed that the model with two descriptors obtained a greater value for correlation coefficient (r = 0.9991, 95% CI_r = [0.9985-0.9994]), comparing with the model with one descriptor (r = 0.9956, 95%CI_r = [0.9929-0.9972]) (Steiger's Z-parameter = 7.016, p < 0.0001).
The hypothesis that there are not significant differences between model from Eq.(3) and previously reported mode (Eq.(1)) was tested by using the Steiger's Z test. The results shown us that the correlation coefficient obtained by Eq.(3) is statistical significant greater comparing with the one obtained by Eq.(1) (Steiger's Z parameter = 2.8, p = $2.6\cdot10^{-3}$).
Figure 1 shown the representation of observed boiling points versus estimated by using the MDF SPR model with two descriptors (Eq.(3)).

Table 1. MDF SPR models for boiling points of alkanes: statistical characteristics

Eq.	95%CI_{inter}	95%$CI_{C\text{-}Desc}$	r^2_{MDF}	F_{MDF}	s_{MDF}	r^2_{loo}	F_{loo}	s_{loo}	r_{sQ}^2
2	[-521.70,-494.22]	[184.22, 192.60]	0.9913	8048‡	3.81	0.9908	7654‡	3.91	0.9571‡
3	[-132.23,-126.16]	[-68.30, -66.60] [4.81, 4.97]	0.9982	19361‡	1.74	0.9980	17837‡	1.82	0.9935‡

95%CI_{inter}, 95%$CI_{C\text{-}Desc}$ = the 95% confidence interval for the intercept, and for the coefficient of descriptor;
r^2_{MDF}, r^2_{loo}= the squared correlation of the MDF SPR model, and for the leave-one-out analysis;
r_{sQ}^2 = semi-quantitative squared correlation coefficient;
F_{MDF}, F_{loo} = the Fisher parameter of the MDF SPR, and leave-one-out regression models;
s_{MDF}, s_{loo} = standard error for the MDF SPR model, and leave-one-out model, respectively; ‡ p < 0.0001

Graphical representation of the observed versus estimated boiling points in training versus test analysis is presented in Figure 2.

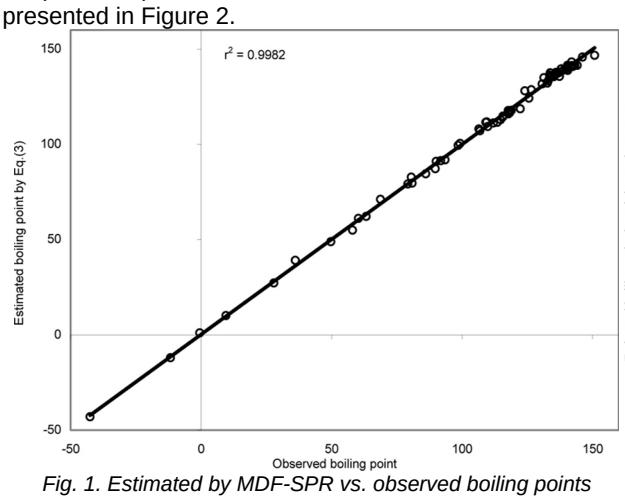


Fig. 1. Estimated by MDF-SPR vs. observed boiling points

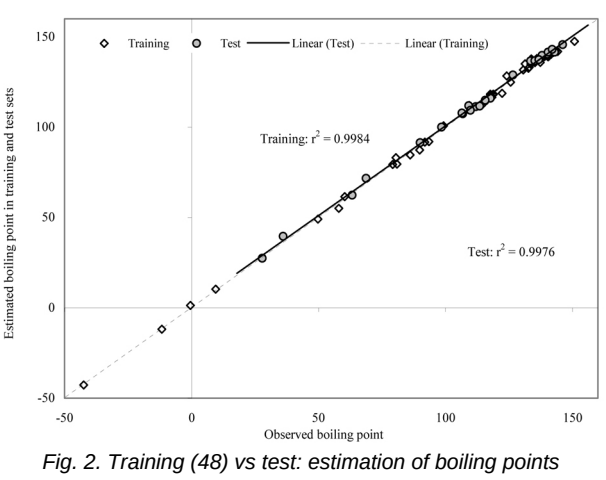


Fig. 2. Training (48) vs test: estimation of boiling points

Performances of model from Eq.(3) in training versus test analysis are presented in Table 1.

Table 1. Validation results of the MDF SPR model with two descriptors - Eq.(3)

Training set			Test set			
No_{tr}	r_{tr}	95%CIr_{tr}	No_{ts}	r_{ts}	95%CIr_{ts}	Zr_{tr}vs r_{ts}
40	0.9992	[0.9987, 0.9994]	33	0.9990	[0.9984, 0.9993]	0.454
41	0.9993	[0.9988, 0.9995]	32	0.9986	[0.9977, 0.9991]	1.406
42	0.9989	[0.9982, 0.9993]	31	0.9993	[0.9988, 0.9995]	0.913
43	0.9988	[0.9980, 0.9992]	30	0.9994	[0.9990, 0.9996]	1.392
44	0.9987	[0.9979, 0.9991]	29	0.9994	[0.9990, 0.9996]	1.543
45	0.9993	[0.9988, 0.9995]	28	0.9986	[0.9977, 0.9991]	1.373
46	0.9991	[0.9985, 0.9994]	27	0.9993	[0.9988, 0.9995]	0.493
47	0.9990	[0.9984, 0.9993]	26	0.9992	[0.9987, 0.9994]	0.434
48	0.9993	[0.9988, 0.9995]	25	0.9979	[0.9966, 0.9986]	2.113†
49	0.9994	[0.9990, 0.9996]	24	0.9985	[0.9976, 0.9990]	1.74†
50	0.9984	[0.9974, 0.9989]	23	0.9995	[0.9992, 0.9996]	2.179†
51	0.9992	[0.9987, 0.9994]	22	0.9987	[0.9979, 0.9991]	0.896
52	0.9991	[0.9985, 0.9994]	21	0.9992	[0.9987, 0.9994]	0.214
53	0.9991	[0.9985, 0.9994]	20	0.9992	[0.9987, 0.9994]	0.21
54	0.9991	[0.9985, 0.9994]	19	0.9993	[0.9988, 0.9995]	0.439
55	0.9990	[0.9984, 0.9993]	18	0.9994	[0.9990, 0.9996]	0.872
56	0.9992	[0.9987, 0.9994]	17	0.9985	[0.9976, 0.9990]	1.047
57	0.9991	[0.9985, 0.9994]	16	0.9992	[0.9987, 0.9994]	0.191
58	0.9991	[0.9985, 0.9994]	15	0.9992	[0.9987, 0.9994]	0.185
59	0.9993	[0.9988, 0.9995]	14	0.9965	[0.9944, 0.9978]	2.442†
60	0.9990	[0.9984, 0.9993]	13	0.9995	[0.9992, 0.9996]	1.011
61	0.9992	[0.9987, 0.9994]	12	0.9962	[0.9939, 0.9976]	2.177†
62	0.9992	[0.9987, 0.9994]	11	0.9920	[0.9872, 0.9949]	3.061†
63	0.9992	[0.9987, 0.9994]	10	0.9971	[0.9953, 0.9981]	1.614

† p < 0.05

CONCLUSIONS

Two MDF-SPR models with good statistical parameters proved to be able to estimate and predict the boiling points of the alkanes with number of atoms that vary from three to nine. Comparing the MDF-SAR models revealed that the model with two descriptors is better.
The descriptors involved into MDF-SAR models were calculated solely from the chemical structure and shown that the boiling point of the studied alkanes depends on topology of compounds and it is related with the group their electronegativity and number of directly bonded hydrogen's.
The internal validation of the MDF-SPR model with two descriptors demonstrates the stability and reliability of the model.
The multi-varied MDF-SPR model can be used in order to predict the boiling point of interest new alkanes (other then the alkanes that were studied) without any experiments and measurements, by using the MDF SPR Predictor application (http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/sar/).

REFERENCES

[1] Toropov, A.; Toropov, A.; Ismailov, T.; Bonchev, D. *J. Mol. Struct. THEOCHEM* 424 (**1998**) 237–247. [2] Jäntschi, L. *Leonardo Electronic Journal of Practices and Technologies* 6 (**2005**) 76-98. [3] Jäntschi, L; Bolboacă, S. *Int. J. Quantum Chem.* 107 (**2007**) 1736-1744. [4] Jäntschi, L; Bolboacă, S. *Int. J. Mol. Sci.* 8 (**2007**) 189-203. [5] Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* 7 (**1991**) 243–272. [6] Steiger, J.H.: Tests for comparing elements of a correlation matrix. Psychol. Bull. 87 (1980) 245-251.

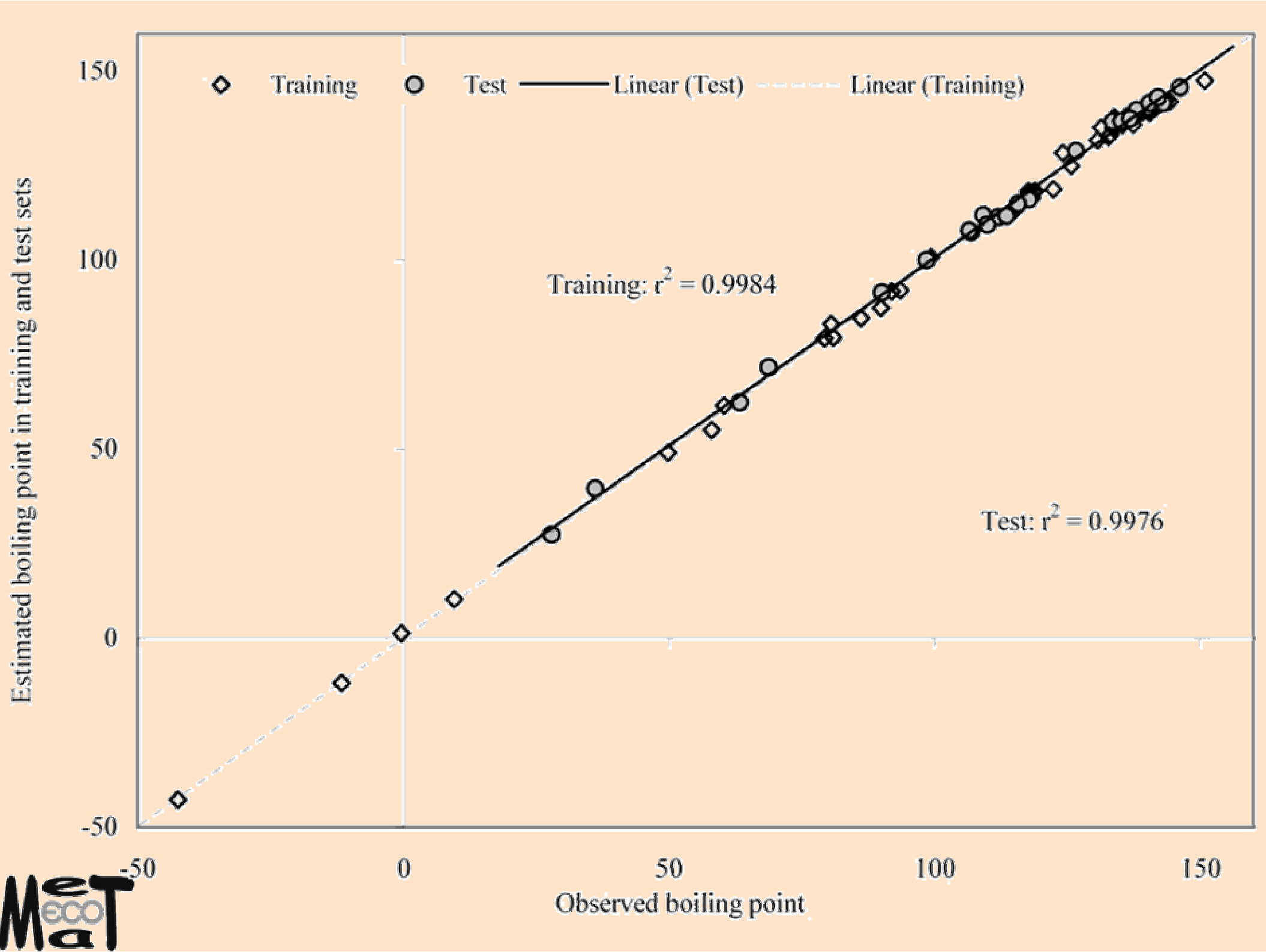
The research was partly supported by UEFISCSU Romania through projects ET36/2005 & ET108/2006.

Modeling the Property of Compounds from Structure: Statistical Methods for Models Validation

Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI

http://sorana.academicdirect.ro http://lori.academicdirect.org

- **Start point**: A previously study of boiling points of a sample of alkanes (Toporov, Toporova, 1998) – it use 3 3D molecular descriptors; $r^2 = 0.9972$
- **Data**: n=78 alkanes; boiling points; structures
- **Method**: MDF-SPR (Jantschi, 2005)
- **Results**: a MDF model with 2 3D molecular descriptors; $r^2 = 0.9982$
- **Discussion**:
  - Leave-one-out analysis $r^2_{loo} = 0.9980$ (!)
  - Training versus test analysis
  - Correlated correlations (Steiger's Z test)


Training
Test
Linear (Test)
Linear (Training)
Training: r2 = 0.9984
Test: r2 = 0.9976
Estimated boiling point in training and test sets
Observed boiling point
150
100
50
0
-50
-50
0
50
100
150

Correlated correlations? - NO

- Analyzing the obtained MDF SPR models revealed that the model with two descriptors obtained a greater value for correlation coefficient (r = 0.9991, 95% CIr = [0.9985-0.9994]), comparing with the model with one descriptor (r = 0.9956, 95%CIr = [0.9929-0.9972]) (Steiger's Z-parameter = 7.016, $p < 0.0001$).
- The hypothesis that there are not significant differences between model with two descriptors and previously reported model was tested by using the Steiger's Z test. The results shown us that the correlation coefficient obtained is statistical significant greater comparing with the one previously obtained (Steiger's Z parameter = 2.8, $p = 2.6 \cdot 10^{-3}$).

Acknowledgements

- UEFISCSU Romania, projects ET36/2005, ET108/2006;
- Thank you for your attention.

MDF ‘ISI’ Refs

- L. J., S.-D. B., Int J Quantum Chem, 2007, 107, 1736;
- L. J., S.-D. B., Int J Mol Sci, 2007, 3, 180;
- S.-D. B., L. J., Int J Mol Sci, 2007, in press;

Specific Support Action
EC-INCO-CT-2005-016414

Workshop on

Ecomaterials and Processes:

Characterization and Metrology

April 19 – 21, 2007, St. Kirik, Plovdiv, Bulgaria

Organised by the

Centre of Competence on
Multifunctional Materials and New Processes with
Environmental Impact (MISSION)
at the Institute of General and Inorganic Chemistry,
Bulgarian Academy of Sciences

http://metecomat.igic.bas.bg