EMPIRICAL MODELING
OF
MASS SPECTRAL FEATURES
BY
MOLECULAR DESCRIPTORS

Kurt Varmuza[1*], Matthias Dehmer[2], Peter Filzmoser[1]

[1] **Vienna University of Technology,**
Department of Statistics and Probability Theory
(and Institute of Chemical Engineering), Laboratory for Chemometrics
Wiedner Hauptstrasse 7/107, A-1040 Vienna, Austria
kvarmuza@email.tuwien.ac.at, www.lcm.tuwien.ac.at/vk/
P.Filzmoser@tuwien.ac.at, www.statistik.tuwien.ac.at/public/filz/

[2] **UMIT, The Health and Life Sciences University,**
Institute for Bioinformatics and Translational Research
Eduard-Wallnöfer Zentrum 1, A-6060 Hall in Tyrol, Austria
matthias.dehmer@umit.at, www.dehmer.org, www.umit.at

* Presenting author

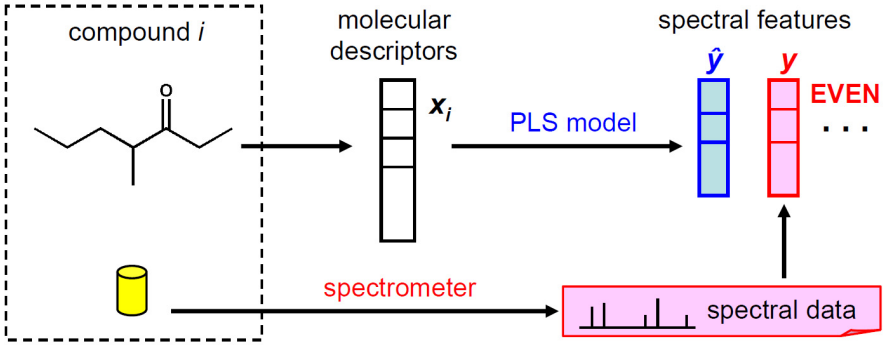
Poster Presentation

Conferentia Chemometrica 2013 (CC 2013)
September 8–11, 2013, Sopron, Hungary

Introduction

Calculation of molecular spectral data from chemical structures is successful in NMR but rather difficult in IR and MS [1].

Preliminary results are presented here for the prediction of a few **mass spectral features** from molecular descriptors (QSPR) for a set of **aliphatic ketones**.



EVEN = $\frac{\text{sum of peak heights at even mass numbers}}{\text{sum of all peak heights}}$ 100 [%]

Example for a mass spectral *feature* (*MS descriptor*) *y*, reflecting ion abundances from competing ion fragmentation pathways - and thus reflecting features of the chemical structure.

Presumptive **application** (besides theoretical interest):

Support of the identification of unknowns by ***systematic structure elucidation*** - especially for compounds not present in spectroscopic databases.

[1] Gasteiger J. (ed.): Handbook of Chemoinformatics - From Data to Knowledge (4 vol.), Wiley-VCH, Weinheim, Germany (2003).

Data

Chemical structures

Chemical structures in Molfile format have been processed by software CORINA [2] resulting in 3D-structures with all H-atoms explicitly given. A set of 150 structures of aliphatic ketones with 5 - 22 C-atoms has been used for QSPR modeling.

The exhaustive set of 194 isomers for $C_{10}H_{20}O$ (aliphatic ketones only) has been generated by software MOLGEN [3].

Molecular descriptors

A set of 741 molecular descriptors (x-variables) has been used, as calculated by software ADRIANA [4].

Mass spectral features

Three mass spectral features (y-variables, calculated by software MassFeatGen [5, 6]) have been modeled by molecular descriptors; they characterize the "shape" of a mass spectrum:

EVEN % of peak intensities at even mass numbers (reflecting competing ion fragmentation pathways)

DUST % of peak intensities at mass numbers ≤ 78 (reflecting the relative abundances of low mass fragments)

IBAS % of base peak intensity (reflecting the distribution of the stability of the ions)

% refers to the sum of all peak intensities

Software

All calculations have been performed within the R environment [7]; some newly developed functions are available [8].

[2] CORINA, version 3.4, software for the generation of high-quality three-dimensional molecular models, Molecular Networks GmbH, Erlangen, Germany, www.molecular-networks.com (2013).
[3] MOLGEN, version 3.5, software for isomer generation, Institute of Mathematics, University of Bayreuth, Germany, www.molgen.de (2000).
[4] ADRIANA, version 2.2.4, software for the encoding of molecular structures, Molecular Networks GmbH, Erlangen, Germany, www.molecular-networks.com (2011).
[5] MassFeatGen, version 1.07, software for calculation of mass spectral features, Laboratory for Chemometrics, Vienna University of Technology, Austria, www.lcm.tuwien.ac.at [Software] (2005).
[6] Demuth W., Karlovits M., Varmuza K.: *Anal. Chim. Acta,* **516**, 75-85 (2004).
[7] R. A language and environment for statistical computing. R Development Core Team, Vienna, Austria, www.r-project.org (2011).
[8] R software and data (test versions) from Laboratory for Chemometrics, Vienna University of Technology, Austria, www.lcm.tuwien.ac.at/R/ (2013).

QSPR Modeling

Strategy

Linear models (separately for each of the 3 y-variables) have been developed with PLS regression. Estimation of the optimum number of PLS components and the prediction performance (for test set objects) - including the variability of these measures - was performed by rdCV (repeated double cross validation). Various variable selection methods were applied and evaluated by rdCV [9].

Performance estimation

The applied rdCV is a resampling method - combining systematics with randomness. rdCV is applicable for calibration and classification for data sets with ca ≥ 25 objects. Optimization of model complexity (optimum no. of PLS components) is separated from the estimation of model performance. rdCV provides estimations of the variabilities of the model complexity and the performance - thus allowing a resonable comparison of models (or variable subsets).

The performance measure used is SEP (standard error of prediction, standard deviation of test-set prediction errors). rdCV typically gives 30 SEP values from repeting the double CV with different random splits into calibration and test sets; these estimations can be represented by a boxplot (or by the mean SEP_{FINAL}).

Variable selection

Three methods have been applied:

- stepwise selection, using BIC as criterion, new R-function [10]
- genetic algorithm (final evaluation by rdCV), R: *subselect/genetic()*
- sequential replacement (final evaluation by rdCV), R: *leaps/regsubsets()*

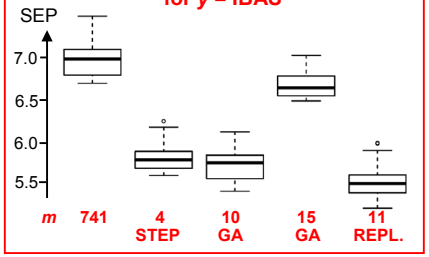
Best results for all 3 y-variables (EVEN, DUST, IBAS): sequential replacement

Comparison of model performance for all descriptors and for best found variable subsets

Modeled *y*	No variable selection		Sequential replacement	
	m	SEP_{FINAL}	*m*	SEP_{FINAL}
EVEN	741	4.5	18	3.1
DUST	741	7.8	16	6.0
IBAS	741	7.0	11	5.5

m, number of variables

Comparison of variable selection for *y* = IBAS



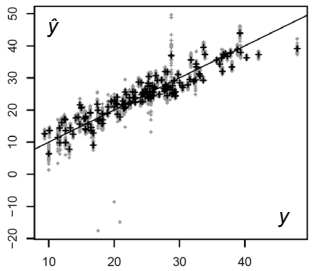
[9] Filzmoser P., Liebmann B., Varmuza K.: *J. Chemom.*, **23**, 160 (2009).
[10] Varmuza K., Filzmoser P., Dehmer M.: Computational and Structural Biotechnology Journal, **5** [6], e201302007, open access (2013).

QSPR Modeling - Results

rdCV results with best variable subsets

n = 150, 30 repetitions, 3 test segments, 4 calibration segments, max. 20 PLS components

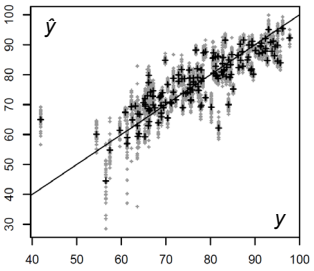
EVEN



m = 18 a_{FINAL} = 17

SEP_{FINAL} = 3.1

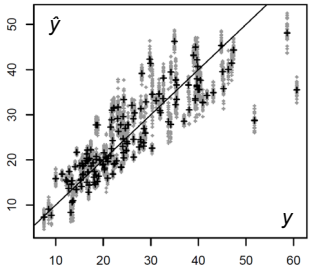
DUST



m = 16 a_{FINAL} = 16

SEP_{FINAL} = 6.0

IBAS



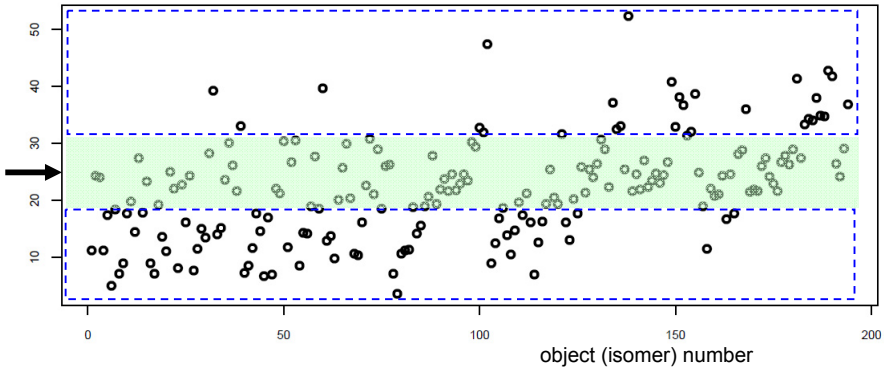
m = 11 a_{FINAL} = 11

SEP_{FINAL} = 5.5

Obviously, these spectral properties are difficult to model

Application of the model for EVEN to *n* = 194 isomers with C_{10}

EVEN predicted



→ Assume, the mass spectrum of an unknown has **EVEN** = 25;
then a **95% confidence interval** of ±2*SEP_{FINAL} = ±2*3.1 = ± 6.2

would **exclude** isomers with EVEN > 31.2 and
with < 18.8

Results - Outlook

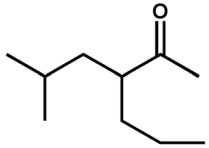
A fictitious example from systematic structure elucidation

The 150 structures used for model creation contain 23 aliphatic ketones with 10 C-atoms (from 194 possible isomers).
Because experimental data (for EVEN, DUST, IBAS) are available for these 23 structures, they may be used for a ***thought experiment:***

Assuming,
- the unknown is an aliphatic ketone,
- and we know the molecular formula (from high resolution mass spec).

How many of the possible isomers can be excluded from the (experimental) mass spectral features EVEN, DUST, and IBAS, using the QSPR models created?

Example



Mass spectral features

	y (exp.)	*ŷ (pred.)*	*error*
EVEN	20.9	19.7	+ 1.2
DUST	84.4	79.7	+ 4.7
IBAS	25.2	27.5	- 2.3

Isomer exclusion

	no. excluded isomers	
EVEN	103	53.1 %
DUST	12	6.2 %
IBAS	17	8.8 %
all 3 ("OR")	115	59.3 %

Summary for all 23 test structures (C_{10} aliphatic ketones)

Using EVEN, DUST and IBAS (with a simple logical OR exclusion), on the average 71 % of the isomers can be excluded.

Cautious conclusion: Perhaps promising to continue ...

Acknowledgments.
We thank J. Gasteiger and C.H. Schwab from Molecular Networks GmbH, www.molecular-networks.com, Erlangen, Germany, for providing software CORINA and ADRIANA. We thank A. Kerber and R. Laue (University of Bayreuth, Germany) for providing the isomer generator software MOLGEN. The work was supported by the Austrian Science Fund (FWF), project P22029-N13.