COSMO data and conventional molecular descriptors compared by redundancy analysis

Bettina Liebmann, Anton Friedl, Kurt Varmuza

Institute of Chemical Engineering, Vienna University of Technology, Getreidemarkt 9, Vienna, Austria.
bettina.liebmann@tuwien.ac.at, www.lcm.tuwien.ac.at, www.vt.tuwien.ac.at

Introduction

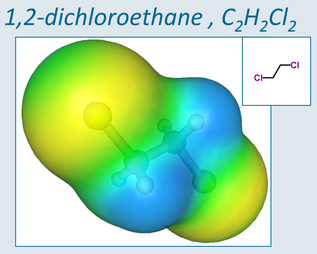
Redundancy analysis (RA) allows comparing the multivariate information content of variable blocks [1]. In quantitative structure-property relationships (QSPR), these variable blocks are groups of molecular descriptors that originate from diverse theoretical concepts. Whenever new types of descriptors are published, it would be helpful to know if they cover similar aspects as already existing descriptors (for a given set of structures). We apply RA to gain insight into the correlations between the blocks of variables, and to assess the extent of redundancy of three COSMO data groups [2] with 28 molecular descriptors groups calculated by software Dragon [3,4]. The asymmetric redundancy index allows not only to quantify the redundant information, but also to spot which of the groups can model the other one. Selected groups are then used as *X* in a QSPR task for the property *y* being the water-octanol partition coefficient, log P.

What is COSMO-RS?

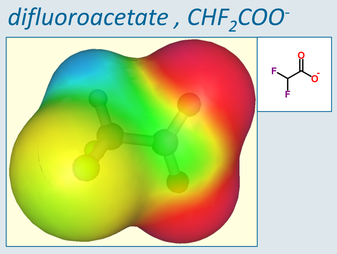
COSMO-RS (**co**nductor-like **s**creening **mo**del for **r**eal **s**olvents) [2] is a quantum chemistry based statistical thermodynamics model for the prediction of thermodynamic properties of fluids and liquid mixtures.

Molecular interactions are calculated from the **screening (polarisation) charge densities (σ)** on the molecular surface, which is defined as the response of an electric conductor to the charge density of the molecule.

1,2-dichloroethane , $C_2H_2Cl_2$



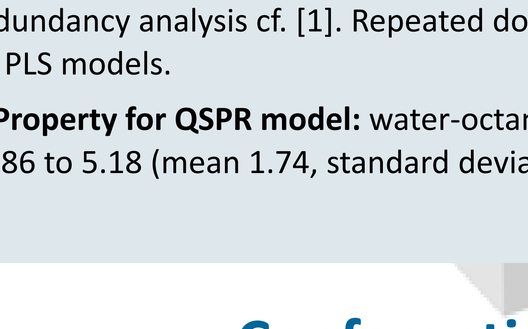
difluoroacetate , CHF_2COO^-



Two sample molecules with their screening charge density σ mapped on the molecular surface ("σ-surface").

blue … negative σ, strong positive polarity
red … positive σ, strong negative polarity
green/yellow … small σ, weak polarity

Each molecule can be characterised by the probability distribution of the screening charge densities σ, called **σ-profile**. The moments derived from the σ-profile are called **σ-moments**, and comprise, e.g., total charge, electrostatic interaction energy, acceptor and donor functions. However, several other moments do not have a simple physical interpretation. The **σ-potential** can be interpreted as the affinity of a solvent for the surface of polarity σ.



***COSMO data for** 1,2-dichloroethane:*

- ***σ-profile** with 61 variables (blue curve, G29)*
- ***σ-potential** with 61 variables (red curve, G30)*
  (61 equidistant σ-values from -0.03 to 0.03 e/A²)
- **σ-moments** with 22 variables (not shown, G31)

Data and Software

- ***n*=131 chemical structures from organic molecules**.

Data format: SMILES (from database [5]) and SDF with approximate 3D structures including all H-atoms (SMILES → 2D SDF [6] → 3D SDF [7])

- ***m*=2691 variables: COSMO data and molecular descriptors, 31 data blocks.**

(G1-G28) Molecular descriptors calculated by software Dragon [4], m_{Dragon} = 2561

(1) Constitutional indices, (2) ring descriptors, (3) topological indices, (4) walk and path counts, (5) connectivity index, (6) information indices, (7) 2D matrix-based descriptors, (8) 2D autocorrelations, (9) Burden eigenvalues, (10) P-VSA-like descriptors, (11) ETA indices, (12) edge adjacency indices, (13) geometrical descriptors, (14) 3D matrix-based descriptors, (15) 3D autocorrelations, (16) RDF descriptors, (17) 3D MoRSE descriptors, (18) WHIM descriptors, (19) GETAWAY descriptors, (20) Randic molecular profiles, (21) functional group counts, (22) atom-centred fragments, (23) atom-type E-state indices, (24) CATS 2D, (25) 2D atom pairs, (26) 3D atom pairs, (27) molecular properties, (28) drug-like indices

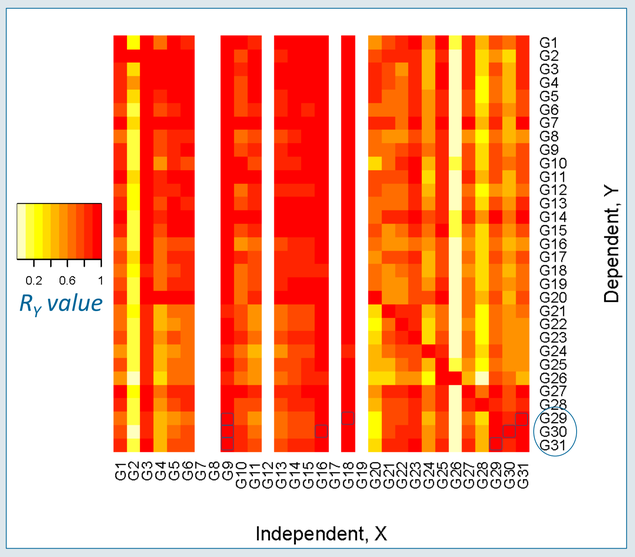
(G29-G31) COSMO data extracted from software COSMOtherm [6], m_{COSMO} = 130

(29) COSMO σ-profile, (30) COSMO σ-potential, (31) COSMO σ-moments

- **Calculations** in R [8], using R-package "chemometrics" and R-code for redundancy analysis cf. [1]. Repeated double cross validation, rdCV [9], used for all PLS models.
- **Property for QSPR model:** water-octanol partition coefficient, log P, values from -0.86 to 5.18 (mean 1.74, standard deviation 1.2) from online database [10].

Correlation Between Variable Groups

The correlations between the 31 groups have been quantified by the **"summed redundancy index R_Y"** (see heatmap below) [1]. Potentially different information of variable groups is indicated by a yellow square, which denotes very low correlation. Group combinations with highest correlations are red. The COSMO groups (G29-G31) have very high correlation (R_Y > 0.94) with "Burden eigenvalues" (G9), "RDF descriptors" (G16), and "WHIM descriptors" (G18). In addition, the redundancy among the COSMO groups is high (R_Y > 0.92).



"Heatmap" of the summed redundancy indices R_Y for 31 variable groups (G1 to G31), values encoded by colour. Note the asymmetry!

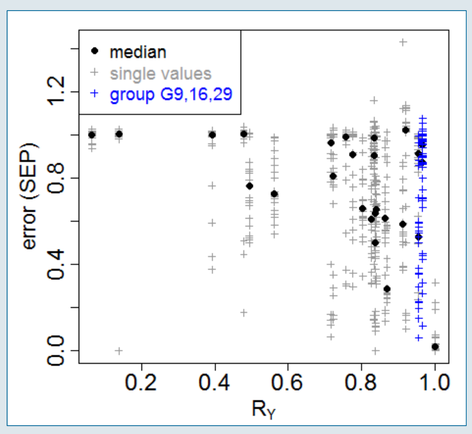
yellow = no or very low correlation
red = high correlation
white = no R_Y values available (for X-groups with m>n, here white columns for G7, 8, 17, 19).

COSMO groups are G29-31. Blue squares indicate group combinations with the three highest R_Y values.

Each of the three Dragon groups have 90-124 variables, and an intrinsic dimensionality (80% variance) of either 5 or 8. The COSMO groups have 21-61 variables, and intrinsic dimensionalities of either 3 or 7.

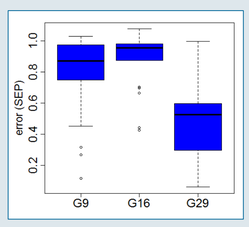
Modelling COSMO Data From Descriptors?

Does a high R_Y imply that a good linear model can be derived for prediction of one group from the other? For each COSMO group, three groups with highest R_Y were selected. For each single COSMO variable used as *y*, a PLS1 model was created with an entire variable group used as *X*. The prediction performance was estimated by rdCV. Result: COSMO data are better modelled by another COSMO group than any single conventional molecular descriptors group (from Dragon).



Model error vs. redundancy plot.

R_Y between COSMO σ-moments (G31) used as Y and all other groups used as X. Test set predicted errors from rdCV for 22 single, autoscaled y-variables of G31 and all other groups used as X in PLS1.



General trend:

For low R_Y, modelling Y from X group is not possible, prediction errors are high. For high R_Y, a good model performance is possible, but not necessarily.

QSPR Model for log P

The best QSPR models for log P are computed from COSMO data, both single COSMO groups and all three together. The good prediction performance could not be reached with Dragon groups; not even with the single Dragon groups G9, 16, 18 that exhibit high R_Y! Despite the redundant information content, COSMO variables obviously contain additional information for modelling the log P. Hence, they can be considered valuable molecular descriptors.

X groups	SEP	*a*
All groups (G1-31)	0.5	8
All Dragon groups (G1-28)	0.6	8
All COSMO groups (G29-31)	**0.4**	**4**
COSMO G29	**0.4**	**8**
COSMO G30	0.5	6
COSMO G31	**0.4**	**4**
Dragon G9	0.9	2
Dragon G16	1.1	1
Dragon G18	1.0	10

SEP, stand. deviation of prediction errors from test sets (rdCV)
a, optimum number of PLS components

References

[1] K. Varmuza, P. Filzmoser, B. Liebmann, M. Dehmer, *Redundancy analysis for characterizing the correlation between groups of variables - applied to molecular descriptors*. Chemom. Intell. Lab. Syst. (2011), doi: 10.1016/j.chemolab.2011.05.013
[2] A. Klamt, *COSMO-RS: From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*, Elsevier Science Ltd., Amsterdam, The Netherlands, 2005.
[3] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, Germany, 2000.
[4] Dragon, Software for molecular descriptor calculation, version 6.0, by R. Todeschini et al., www.talete.mi.it
[5] COSMOthermX, version C2.1, release 01.10; COSMOlogic GmbH & Co. KG, Leverkusen, Germany, 2009.
[6] The Open Babel Package, version 2.3.0, http://openbabel.org
[7] Corina software, Molecular Networks GmbH Computerchemie, www.mol-net.de, Erlangen, Germany, 2004.
[8] R. A language and environment for statistical computing, version 2.12.2. R Development Core Team, Vienna, Austria, 2011. www.r-project.org
[9] P. Filzmoser, B. Liebmann, K. Varmuza, *Repeated Double Cross Validation*, J. Chemom., 23, 160-171 (2009).
[10] Physical Properties Database *PhysProp*, www.srcinc.com/what-we-do/databaseforms.aspx?id=386

Acknowledgements

We thank Prof. Peter Filzmoser for contributing R code, and Dr. Andreas Klamt, COSMOlogic for providing COSMOthermX software.

Conferentia Chemometrica 2011 – Sümeg, Hungary, 18.-21.9.2011