chem-bla-ics

From the archives: my ICCS 2005 poster

Published June 25, 2011

Citation
Willighagen, E. (2011, June 25). From the archives: my ICCS 2005 poster. *Chem-bla-ics*. https://doi.org/10.59350/n4hbf-t3t23

Keywords
Iccs

Abstract
Julio and Gert placed their ICCS 2011 work online, and today I was going through old CDs (see From the archives: Chemical Web, and the CDK in 2004 and Chiral Molecules: how cool is the SEM picture?). I also ran into my ICCS 2005 poster, and because that too was before I started blogging, I never posted it online.

Julio and Gert placed their ICCS 2011 work online, and today I was going through old CDs (see From the archives: Chemical Web, and the CDK in 2004 and Chiral Molecules: how cool is the SEM picture?). I also ran into my ICCS 2005 poster, and because that too was before I started blogging, I never posted it online. So, here it is, based on my thesis :

On the use of ^{1}H and ^{13}C NMR spectra as QSAR descriptors

E.L. Willighagen, R. Wehrens, and L.M.C. Buydens

Institute for Molecules and Materials
Radboud University Nijmegen

e.willighagen@science.ru.nl

Introduction

Quantitative Structure Activity Relationship (QSAR) models correlate molecular structures with biological and chemical activities.

Spectra have been suggested as descriptor of the molecular structures, but the performance of spectra-based QSAR models has not been thoroughly tested.

This poster presents QSAR models based on ^{1}H and ^{13}C NMR spectra and compares this with models build from theoretical molecular descriptors.

Data Sets

Three data sets are discussed:

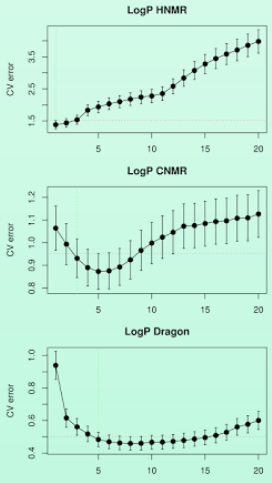
name	# compounds	activity	ref.
WS	431	water solubility	[1]
BP	277	boiling point	[2]
LogP	154	LogP	[3]

For each data set ^{1}H and ^{13}C NMR spectra are simulated using ACD/Labs NMR Predictor. Theoretical molecular descriptors are calculated with Dragon and a subset is randomly chosen. All three descriptor sets contain 220 variables.

Methods

Partial Least Squares (PLS) was used to make the regression models. leave-one-out cross validation (LOO-CV) was used to pick the right number of latent variables (LV's).

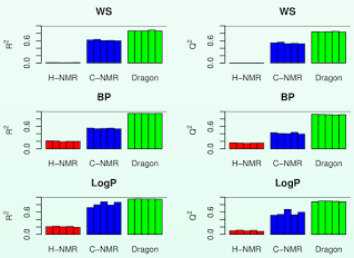
The vertical dotted line indicates the selected number of latent variables. Whiskers indicate ± 1 standard deviation in the cross validation error:



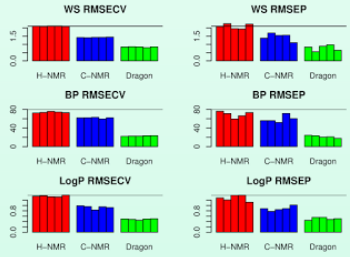
For each model type five randomly chosen independent test sets were used.

Results

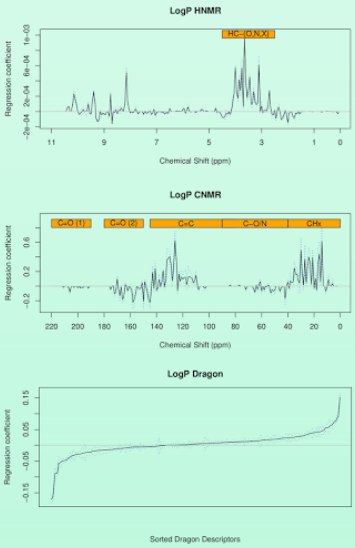
The internal performance statistics R^2 and Q^2 for the three data sets show that ^{1}H NMR does not yield acceptable models. While ^{13}C NMR models are acceptable, Dragon descriptors are clearly better:



This is confirmed by the root mean square errors for the LOO-CV (RMSECV) and for the predictions of the test set (RMSEP). The horizontal line indicates the error of a $y_{pred} = \bar{y}$ model; RMSE values should be well below this limit:

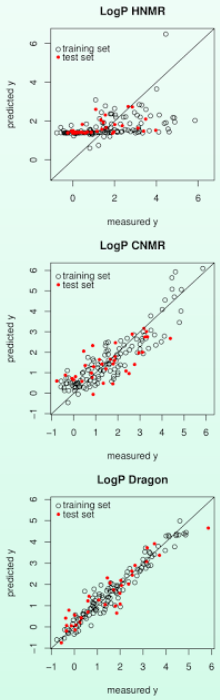


The regression vector of the PLS models for ^{1}H NMR shows much less structure than the ^{13}C NMR. In blue are the ± 1 standard deviations of the five models:



LogP Prediction

Visual inspection of the $y_{predicted}$ vs. $y_{measured}$ plots confirms that Dragon-based models give more accurate predictions than ^{13}C NMR-based models for both the training set (black) and the independent test set (red):



Conclusions

- ^{1}H NMR spectra do not yield good PLS regression models.
- ^{13}C NMR spectra yield acceptable PLS regression models, but are inferior to models based on theoretical molecular descriptors

References

[1] A. Yan and J. Gasteiger. Prediction of Aqueous Solubility of Organic Compounds based on a 3D Structure Representation. *J.Chem.Inf.Comput.Sci.*, 43:429–434, 2003.

[2] E.S. Goll and P.C. Jurs. Prediction of the Normal Boiling Points of Organic Compounds from Molecular Structures with a Computational Neural Network Model. *J.Chem.Inf.Comput.Sci.*, 39:974–983, 1999.

[3] L.K. Schnackenberg and R.D. Beger. Whole-Molecule Calculation of Log P Based on Molar Volume, Hydrogen Bonds, and Simulated ^{13}C NMR Spectra. *J.Chem.Inf.Model.*, 45:360–365, 2005.