Abstract

Computational and Structural Biotechnology Journal, **5**, [6] e201302007, 1-10 (2013)
http://journals.sfu.ca/rncsb/index.php/csbj/article/view/csbj.201302007/224
http://dx.doi.org/10.5936/csbj.201302007

Varmuza K., Filzmoser P., Dehmer M.:

Multivariate linear QSPR/QSAR models: Rigorous evaluation of variable selection for PLS.

Basic chemometric methods for making empirical regression models for QSPR/QSAR are briefly described from a user's point of view. Emphasis is given to PLS regression, simple variable selection and a careful and cautious evaluation of the performance of PLS models by repeated double cross validation (rdCV).

A demonstration example is worked out for QSPR models that predict gas chromatographic retention indices (values between 197 and 504 units) of 209 polycyclic aromatic compounds (PAC) from molecular descriptors generated by *Dragon* software. Most favorable models were obtained from data sets containing also descriptors from 3D structures with all H-atoms (computed by *Corina* software), using stepwise variable selection (reducing 2688 descriptors to a subset of 22).

The final QSPR model has typical prediction errors for the retention index of $\pm$12 units (95% tolerance interval, for test set objects). Programs and data are provided as supplementary material for the open source *R* software environment.