SAR-*caddle.*®

(available in early 2013)

SAR-caddle.® is an entirely web-based SAR-program that offers especially robust interpolation methods for building and applying Structure-Activity (SAR) and Structure-Property Relationships (SPR).

Web Based:

There is no need to install *SAR-caddle.*® anywhere other than on the central server. This means that *SAR-caddle.*® can be used on any desktop computer, laptop or even smart phone or tablet PC that can access the server. The advantages of this architecture are:

- Ease of installation and low maintenance
- High-performance compute modules run on the server, which may be highly parallel or use graphics processors to ensure short turnaround. Compute modules can use high-performance libraries and other features generally not available on desktop machines.
- The compute resources of the server can be coordinated optimally by the*SAR-caddle.*® server
- No license problems: *SAR-caddle.*® is available for all users that can access the *SAR-caddle.*® URL

Easy Data Preparation:

SAR-caddle.® uses either standard MicroSoft Excel® .xls or .xlsx files or tab- or comma-separated tables as input.

Data Analysis:

The initial data analysis performed by *SAR-caddle.*® ensures data integrity and suitability for SAR analyses. The checks performed include:

- Checking for missing or inappropriate data
- Detecting and eliminating identical data points
- Checking for data distribution and suggesting whether to use raw or log10 data as input
- Detecting and optionally eliminating data points that are mutually inconsistent
- Calculating the correlation matrix and optionally eliminating one of each pair of highly correlated descriptors
- Performing a principal-component analysis to determine the inherent dimensionality of the descriptors. The principal components are thus also available for model building, visualization or analysis

Model Building:

SAR-caddle.® uses a variety of very robust techniques to build statistically valid, reliable and predictive models. A unique "traffic light" system indicates the suitability of data and descriptors for inclusion in the model (either in their original form or as log values). The techniques available include:

- ***Bagging stepwise regression***: The dataset is divided into a large number overlapping test/training set combinations. Each is subjected to a stepwise linear regression using the newly defined[1] critical F-values that take descriptor bias into account. This extremely conservative regression procedure avoids overtraining (a common feature of standard multiple regression programs) and also provides a realistic estimate of the likely prediction error.[2]
- ***Partial Least Squares (PLS)***: The *SAR-caddle.*® implementation of PLS also uses a committee-machine strategy[3] that provides error estimates for the model predictions.
- ***Shepard Interpolation*** Shepard interpolation is a simple and robust technique to estimate the value of a query sample by distance-weighted interpolation.

Data Visualization:

Interactive 3D-visualization of the data is available (without installing additional software or plugins) using technology developed by Molcad GmbH. High-quality diagrams can be rotated and zoomed interactively for an easier understanding of the underlying structure of the data. Currently, *SAR-caddle.*® includes two types of visualization.

- ***2D scatter plots of descriptor correlations***: By clicking the correlation matrix specific correlations between two descriptors will be depicted as 2D scatter plot including detailed information for selected data points. This will ease to detect data correlations like log10.
- ***Color-coded 3D-scatter plots***: Can use principal components to display data, making clustering etc. immediately visible.

Who Should Use SAR-caddle.®?

SAR-caddle.® is intended to allow non-specialist users to extract predictive, robust SAR models from their data. In the "comfort" mode, *SAR-caddle.*® makes all the necessary decisions necessary, performs all the model building that it thinks advisable and reports the results. *SAR-caddle.*® will not find a model if the data do not provide one and will report accordingly. *SAR-caddle.*® is particularly suitable for such applications because it works with standard Excel® .xls files as input.

[1] *Sharpening the toolbox of computational chemistry: A new approximation of critical F-values for multiple linear regression, C. Kramer, C. S. Tautermann, D. J. Livingstone, D. W. Salt, D. C. Whitley, B. Beck and T. Clark, J. Chem. Inf. Model. ,* ***2009****, 49, 28-34.*

[2] *Conformation-Dependent QSPR-Models: logP$_{OW}$*, M. Muehlbacher, A. El Kerdawy, C. Kramer, B. D. Hudson and T. Clark, *J. Chem. Inf. Model.*, **2011**, *51*, 2408-2416

[3] *QM/NN QSPR Models with Error Estimation: Vapor Pressure and logP , B. Beck, A. Breindl and T. Clark,J. Chem. Inf. Comput. Sci.,***2000***,40*, 1046-1051.