SAR-caddle® User's Manual

A web-based tool for creating and applying Structure-Activity Relationship models

Index of contents:

1. Introduction

What is the purpose of *SAR-caddle*®?

SAR-caddle® is an entirely web-based program that offers especially robust interpolation methods for building Structure-Activity (SAR) and Structure-Property Relationships (SPR) , and then applies the models to make predictions. It allows previous experimental work to be used to make fast predictions to help guide future research, saving both time and money. It is able to predict any arbitrary property, such as logP (water-octanol partition coefficient), melting point, friction or oil separation, etc., provided that sufficient data is available.

Who should use *SAR-caddle*®?

SAR-caddle® is designed to be used by scientists and engineers who are not experts in mathematics or statistics (such as experimental chemists, biochemists and chemical engineers) to extract predictive, robust models from their data. However SAR specialists with the prerequisite experience in this area will also benefit from the state-of-the-art methods in *SAR-caddle*®

How does *SAR-caddle*® work?

SAR-caddle® takes a file of empirically measured properties and uses these to construct models to reproduce an experimental property (Structure-Activity Relationships (SAR)) by applying techniques such as linear regression, partial least squares or Shepard Interpolation. SAR-caddle® makes all the necessary decisions, 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 support one and will report accordingly. Moreover, SAR-caddle® includes automated built-in safeguards to evaluate and report on the predictivity and applicability of the model created. SAR-caddle® is particularly suitable for such applications because it works with standard Excel® .xls or .xlsx files or a comma- or tab-separated ASCII text files as input.

The descriptors required for SAR-caddle can either be derived from previous experimental work or be created automatically using ParaSurf™ or calculated using other descriptor-generating programs such as Dragon or Codessa.

How is *SAR-caddle*® installed?

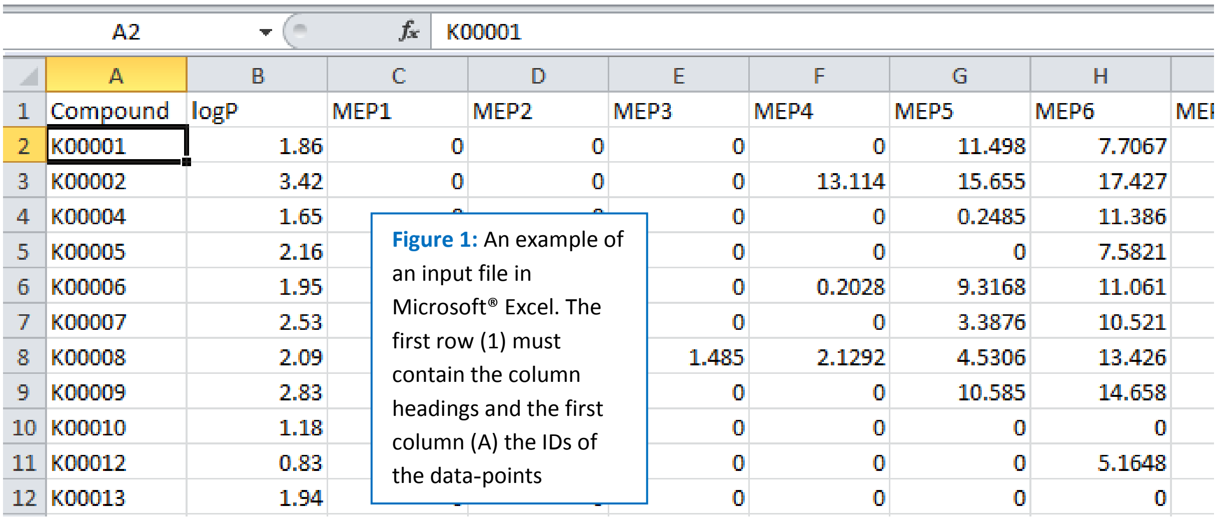
SAR-caddle® is a web-site server for processing data and the users access it through a conventional browser window. It can be installed locally on an isolated computer or within an intranet. There is no need to install SAR-caddle® anywhere other than on the central server. It is then available to all users on a variety of different platforms. All that is required is a suitable browser (see http://www.ceposinsilico.de/products/caddle.htm for a list of compatible browsers). 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 computational resources of the server can be coordinated optimally by the SAR-caddle® server
- No licensing or installation issues: SAR-caddle® is available for all users that can access the SAR-caddle® URL

2. Input data for SAR-caddle®

Input to SAR-caddle® is a simple Microsoft® Excel file (.xls or .xlsx) or a comma- or tab-separated ASCII text file. Excel files can also be generated from Libre office/Open Office using the "*save as Microsoft Excel 97*" option, from Mac using the "*save as xlsx*" or "*save as xls 97-04*" options. The first row of data should contain the column names and the first column should contain the IDs of the data-points (e.g. compound names, registry numbers, etc.). Figure 1 shows a section of an example file (logP_100.xls in the example data collection).



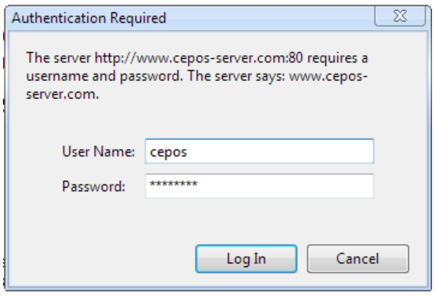
	A	B	C	D	E	F	G	H	MEI
1	Compound	logP	MEP1	MEP2	MEP3	MEP4	MEP5	MEP6	MEI
2	K00001	1.86	0	0	0	0	11.498	7.7067	
3	K00002	3.42	0	0	0	13.114	15.655	17.427	
4	K00004	1.65			0	0	0.2485	11.386	
5	K00005	2.16			0	0	0	7.5821	
6	K00006	1.95			0	0.2028	9.3168	11.061	
7	K00007	2.53			0	0	3.3876	10.521	
8	K00008	2.09			1.485	2.1292	4.5306	13.426	
9	K00009	2.83			0	0	10.585	14.658	
10	K00010	1.18			0	0	0	0	
11	K00012	0.83			0	0	0	5.1648	
12	K00013	1.94			0	0	0	0	

Figure 1: An example of an input file in Microsoft® Excel. The first row (1) must contain the column headings and the first column (A) the IDs of the data-points

During the SAR-caddle® workflow, one of the properties (columns B to H) is selected as the predictor variable to be modeled, and a subset of the remaining properties is chosen to provide the model descriptors . In this case, LogP will be modeled and the calculated binned molecular electrostatic potential (ParaSurf ™, MEP1 to MEP6) are potential descriptors.

3. Getting started

If your system manager has set up SAR-caddle® to require user names and passwords, your first view of SAR-caddle® when you access the URL provided by your system administrator will be the login page shown in Figure 2, which is self-explanatory.



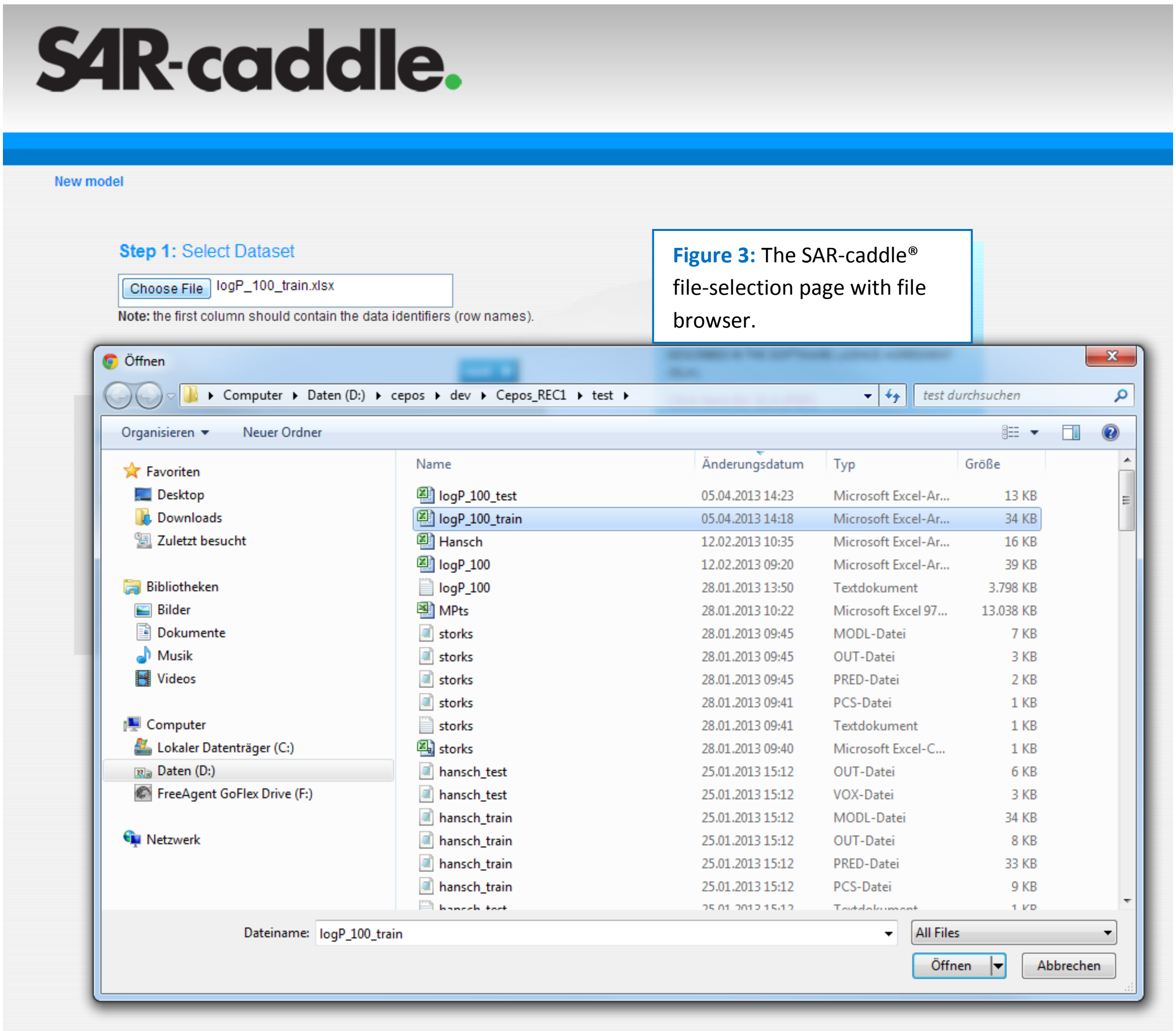
The server http://www.cepos-server.com:80 requires a username and password. The server says: www.cepos-server.com.

User Name: cepos

Password: ********

Figure 2: The SAR-caddle® login box.

Once you have logged into the system, SAR-caddle® allows you to enter the name of an input file or to choose one using the file-browser (see Figure 3).



Step 1: Select Dataset

Choose File logP_100_train.xlsx

Note: the first column should contain the data identifiers (row names).

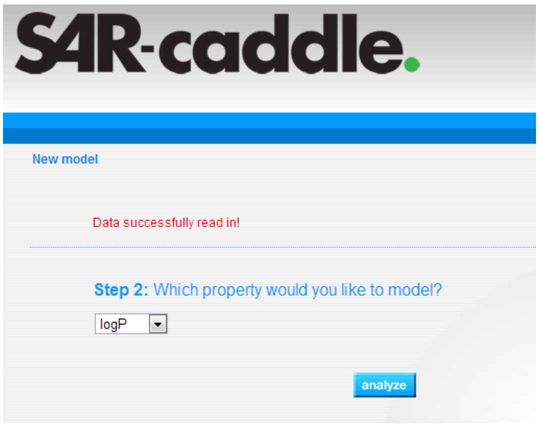
Name	Änderungsdatum	Typ	Größe
logP_100_test	05.04.2013 14:23	Microsoft Excel-Ar...	13 KB
logP_100_train	05.04.2013 14:18	Microsoft Excel-Ar...	34 KB
Hansch	12.02.2013 10:35	Microsoft Excel-Ar...	16 KB
logP_100	12.02.2013 09:20	Microsoft Excel-Ar...	39 KB
logP_100	28.01.2013 13:50	Textdokument	3.798 KB
MPts	28.01.2013 10:22	Microsoft Excel 97...	13.038 KB
storks	28.01.2013 09:45	MODL-Datei	7 KB
storks	28.01.2013 09:45	OUT-Datei	3 KB
storks	28.01.2013 09:45	PRED-Datei	2 KB
storks	28.01.2013 09:41	PCS-Datei	1 KB
storks	28.01.2013 09:41	Textdokument	1 KB
storks	28.01.2013 09:40	Microsoft Excel-C...	1 KB
hansch_test	25.01.2013 15:12	OUT-Datei	6 KB
hansch_test	25.01.2013 15:12	VOX-Datei	3 KB
hansch_train	25.01.2013 15:12	MODL-Datei	34 KB
hansch_train	25.01.2013 15:12	OUT-Datei	8 KB
hansch_train	25.01.2013 15:12	PRED-Datei	33 KB
hansch_train	25.01.2013 15:12	PCS-Datei	9 KB
hansch_test	25.01.2013 15:12	Textdokument	1 KB

Dateiname: logP_100_train

All Files

Figure 3: The SAR-caddle® file-selection page with file browser.

Clicking on next ▶ instructs SAR-caddle® to read in the data. If the format of the data file is correct, SAR-caddle® will indicate this with the message "Data successfully read in!" and move on to the next page, which allows you to select the data that you would like to model in a pull-down menu, as shown in Figure 4.



Data successfully read in!

Step 2: Which property would you like to model?

logP

Figure 4: The SAR-caddle® data-selection page. The file browser features a pull-down menu to select the data to be modeled.

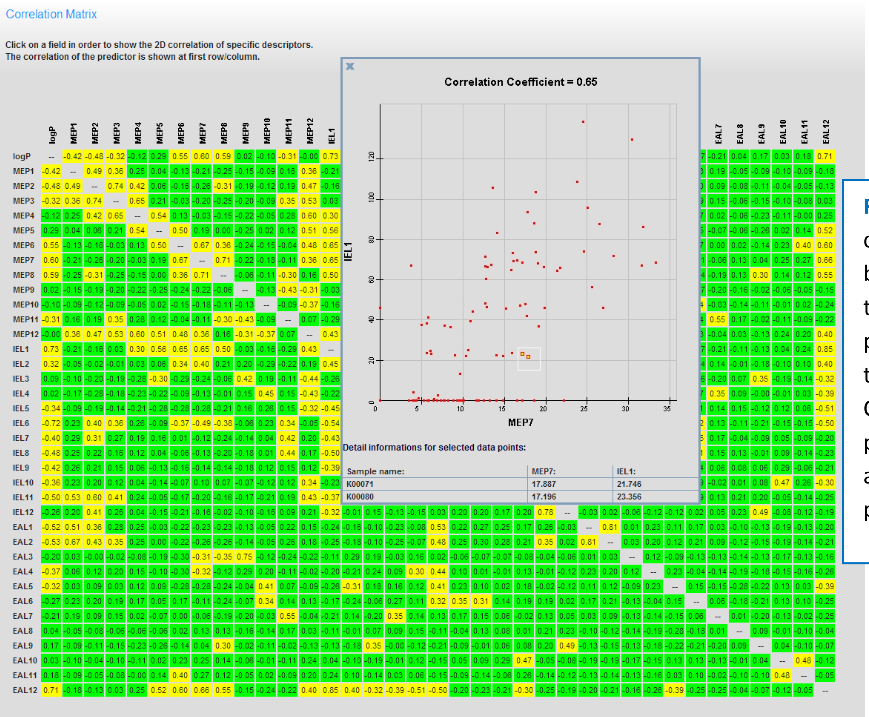
In this example, we select “logP” and proceed by clicking the analyze button.

3.1. Error report

Error Message	**Please check**
The following data Columns were ignored	your input file for empty cells or strings in the columns mentioned.

4. Data Analysis

SAR-caddle® then provides an initial analysis of your data. This provides two important pieces of information. First, the correlation matrix (shown in Figure 5) provides a simple color-coded matrix of the correlation coefficients (R) between all the variables (columns) in the input file. Red indicates highly correlated, yellow moderately correlated and green poorly correlated variables. Clicking on an entry in the matrix displays a plot of the two corresponding columns of data against each other, as shown in Figure 5. Calculating the correlation matrix is important to optionally eliminating one of each pair of highly correlated descriptors.



Correlation Matrix

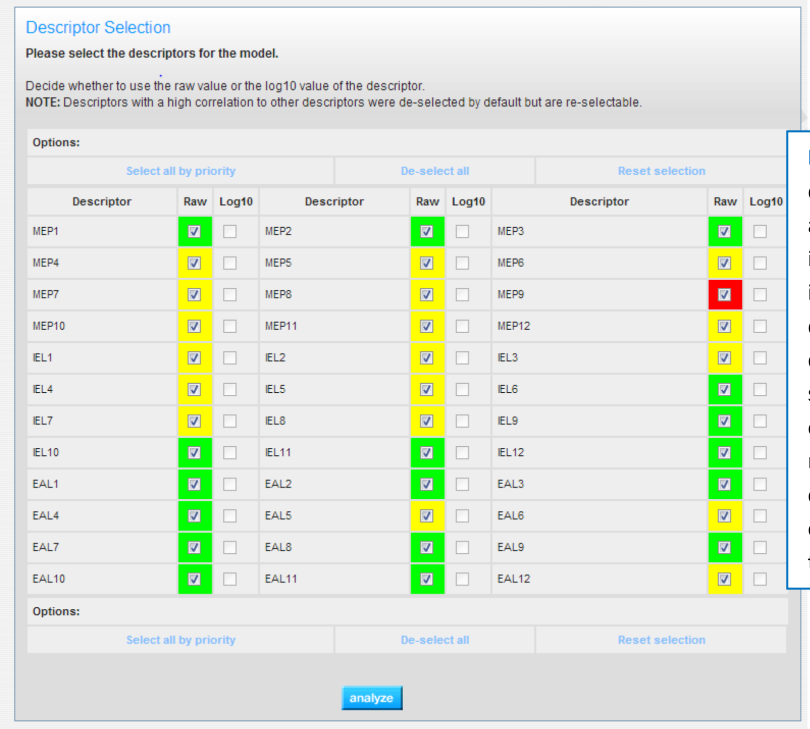
Click on a field in order to show the 2D correlation of specific descriptors.
The correlation of the predictor is shown at first row/column.

Detail informations for selected data points:

Sample name:	MEP7:	IEL1:
K00071	17.887	21.746
K00080	17.196	23.356

Figure 5: The SAR-caddle® correlation matrix. Clicking on a box in the matrix (in this case that between birth rate and population) provides a plot of the two properties as shown. Clicking on data points in the plot provides details of the data and those of the surrounding points.

The second important analysis is that shown in Figure 6. SAR-caddle® investigates the relationships between the individual columns of data and the distribution of the data within the columns in order to recommend which columns to include in the subsequent analysis.



Descriptor Selection

Please select the descriptors for the model.

Decide whether to use the raw value or the log10 value of the descriptor.
NOTE: Descriptors with a high correlation to other descriptors were de-selected by default but are re-selectable.

Options:

Descriptor	Raw	Log10	Descriptor	Raw	Log10	Descriptor	Raw	Log10
MEP1	🟩☑	☐	MEP2	🟩☑	☐	MEP3	🟩☑	☐
MEP4	🟨☑	☐	MEP5	🟨☑	☐	MEP6	🟨☑	☐
MEP7	🟨☑	☐	MEP8	🟨☑	☐	MEP9	🟥☑	☐
MEP10	🟨☑	☐	MEP11	🟨☑	☐	MEP12	🟨☑	☐
IEL1	🟨☑	☐	IEL2	🟨☑	☐	IEL3	🟨☑	☐
IEL4	🟨☑	☐	IEL5	🟨☑	☐	IEL6	🟩☑	☐
IEL7	🟨☑	☐	IEL8	🟨☑	☐	IEL9	🟩☑	☐
IEL10	🟩☑	☐	IEL11	🟩☑	☐	IEL12	🟩☑	☐
EAL1	🟩☑	☐	EAL2	🟩☑	☐	EAL3	🟩☑	☐
EAL4	🟩☑	☐	EAL5	🟨☑	☐	EAL6	🟨☑	☐
EAL7	🟩☑	☐	EAL8	🟩☑	☐	EAL9	🟩☑	☐
EAL10	🟩☑	☐	EAL11	🟩☑	☐	EAL12	🟨☑	☐

Options:

Figure 6: The SAR-caddle® analysis of the input data. In this example, none of the data columns are strongly correlated and so they are all included in the analysis. The color coding indicates how closely to normal the data are distributed. Green indicates an essentially normal distribution, yellow skewed and red either very skewed or bimodal. The column "MEP9" contains only positive non-zero values and is therefore not really suitable as a descriptor. In this case, SAR-caddle® has included it. This choice can be overridden by clicking on the red box to remove the tick..

In this example, we choose to override the recommendation that the model be built using log (MEP9) by clicking the corresponding "raw" box, which deactivates the "Log10" selection (Fig. 6 inset). Clicking the analyze button requests that SAR-caddle® proceed to the model-building step. In this example, none of the descriptors (columns) are highly correlated. If SAR-caddle® finds two highly correlated descriptors, it removes one from the descriptor list before moving on to the model-building analysis. This is because highly correlated descriptors can be combined in arbitrary proportions to give the same result. This renders the regression results ambiguous and does not improve the quality of the fit. Similarly, descriptors that contain no information are also removed. The automatic choices made by the program can be overridden by the user.

5. Standard SAR-caddle® models

The standard SAR-caddle® modeling analysis provides five different analyses of the data used to build the model and their relationship to the target property (in this case logP). The first is the correlation matrix, which is exactly analogous to that shown in Figure 5 (and has the same functionality) but only includes the descriptors (data columns) used to model the data.

5.1. Principal Components Analysis (PCA)

Principal components analysis (PCA) is a technique used for data reduction and analysis in which the interrelationships of the data columns are investigated. Briefly, principal components are eigenvectors of the correlation matrix between variables (data columns). Their associated eigenvalues allow us to judge the dimensionality of the dataset (i.e. how many data columns do we

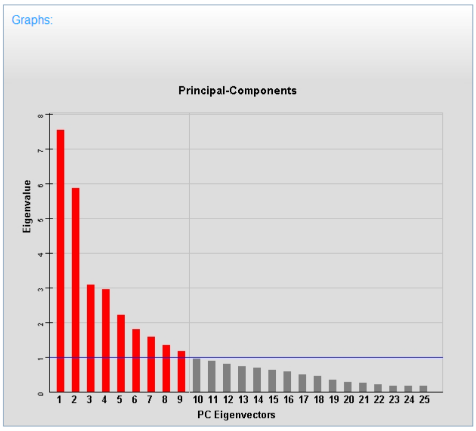
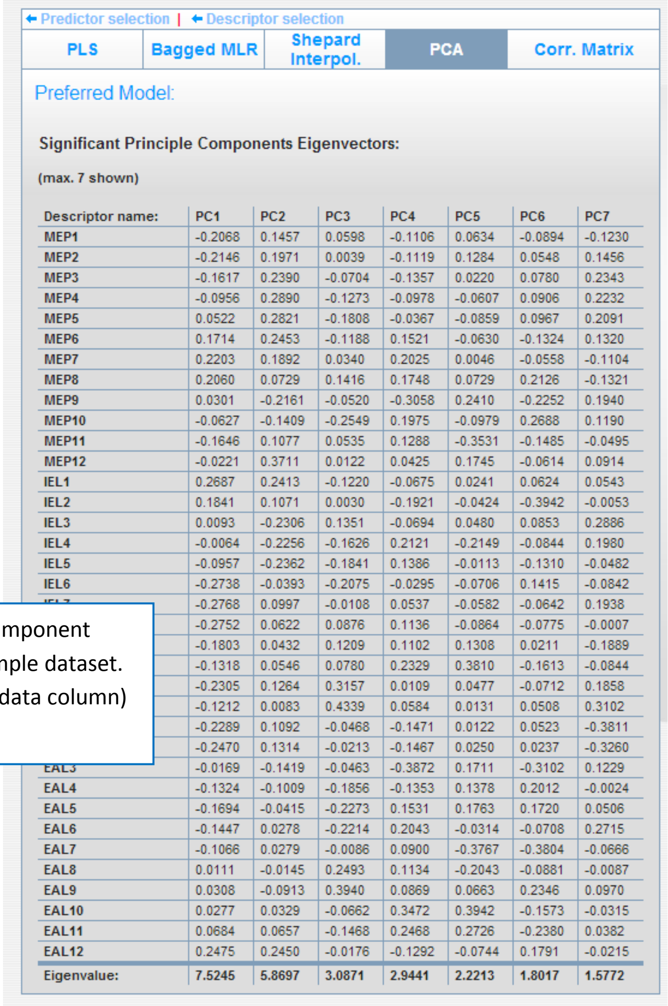


Figure 7: The Eigenvalue plot from the SAR-caddle® PCA-analysis. In this example, the first nine principal components are significant.



Preferred Model:

Significant Principle Components Eigenvectors:

(max. 7 shown)

Descriptor name:	PC1	PC2	PC3	PC4	PC5	PC6	PC7
MEP1	-0.2068	0.1457	0.0598	-0.1106	0.0634	-0.0894	-0.1230
MEP2	-0.2146	0.1971	0.0039	-0.1119	0.1284	0.0548	0.1456
MEP3	-0.1617	0.2390	-0.0704	-0.1357	0.0220	0.0780	0.2343
MEP4	-0.0956	0.2890	-0.1273	-0.0978	-0.0607	0.0906	0.2232
MEP5	0.0522	0.2821	-0.1808	-0.0367	-0.0859	0.0967	0.2091
MEP6	0.1714	0.2453	-0.1188	0.1521	-0.0630	-0.1324	0.1320
MEP7	0.2203	0.1892	0.0340	0.2025	0.0046	-0.0558	-0.1104
MEP8	0.2060	0.0729	0.1416	0.1748	0.0729	0.2126	-0.1321
MEP9	0.0301	-0.2161	-0.0520	-0.3058	0.2410	-0.2252	0.1940
MEP10	-0.0627	-0.1409	-0.2549	0.1975	-0.0979	0.2688	0.1190
MEP11	-0.1646	0.1077	0.0535	0.1288	-0.3531	-0.1485	-0.0495
MEP12	-0.0221	0.3711	0.0122	0.0425	0.1745	-0.0614	0.0914
IEL1	0.2687	0.2413	-0.1220	-0.0675	0.0241	0.0624	0.0543
IEL2	0.1841	0.1071	0.0030	-0.1921	-0.0424	-0.3942	-0.0053
IEL3	0.0093	-0.2306	0.1351	-0.0694	0.0480	0.0853	0.2886
IEL4	-0.0064	-0.2256	-0.1626	0.2121	-0.2149	-0.0844	0.1980
IEL5	-0.0957	-0.2362	-0.1841	0.1386	-0.0113	-0.1310	-0.0482
IEL6	-0.2738	-0.0393	-0.2075	-0.0295	-0.0706	0.1415	-0.0842
IEL7	-0.2768	0.0997	-0.0108	0.0537	-0.0582	-0.0642	0.1938
	-0.2752	0.0622	0.0876	0.1136	-0.0864	-0.0775	-0.0007
	-0.1803	0.0432	0.1209	0.1102	0.1308	0.0211	-0.1889
	-0.1318	0.0546	0.0780	0.2329	0.3810	-0.1613	-0.0844
	-0.2305	0.1264	0.3157	0.0109	0.0477	-0.0712	0.1858
	-0.1212	0.0083	0.4339	0.0584	0.0131	0.0508	0.3102
	-0.2289	0.1092	-0.0468	-0.1471	0.0122	0.0523	-0.3811
	-0.2470	0.1314	-0.0213	-0.1467	0.0250	0.0237	-0.3260
EAL3	-0.0169	-0.1419	-0.0463	-0.3872	0.1711	-0.3102	0.1229
EAL4	-0.1324	-0.1009	-0.1856	-0.1353	0.1378	0.2012	-0.0024
EAL5	-0.1694	-0.0415	-0.2273	0.1531	0.1763	0.1720	0.0506
EAL6	-0.1447	0.0278	-0.2214	0.2043	-0.0314	-0.0708	0.2715
EAL7	-0.1066	0.0279	-0.0086	0.0900	-0.3767	-0.3804	-0.0666
EAL8	0.0111	-0.0145	0.2493	0.1134	-0.2043	-0.0881	-0.0087
EAL9	0.0308	-0.0913	0.3940	0.0869	0.0663	0.2346	0.0970
EAL10	0.0277	0.0329	-0.0662	0.3472	0.3942	-0.1573	-0.0315
EAL11	0.0684	0.0657	-0.1468	0.2468	0.2726	-0.2380	0.0382
EAL12	0.2475	0.2450	-0.0176	-0.1292	-0.0744	0.1791	-0.0215
Eigenvalue:	7.5245	5.8697	3.0871	2.9441	2.2213	1.8017	1.5772

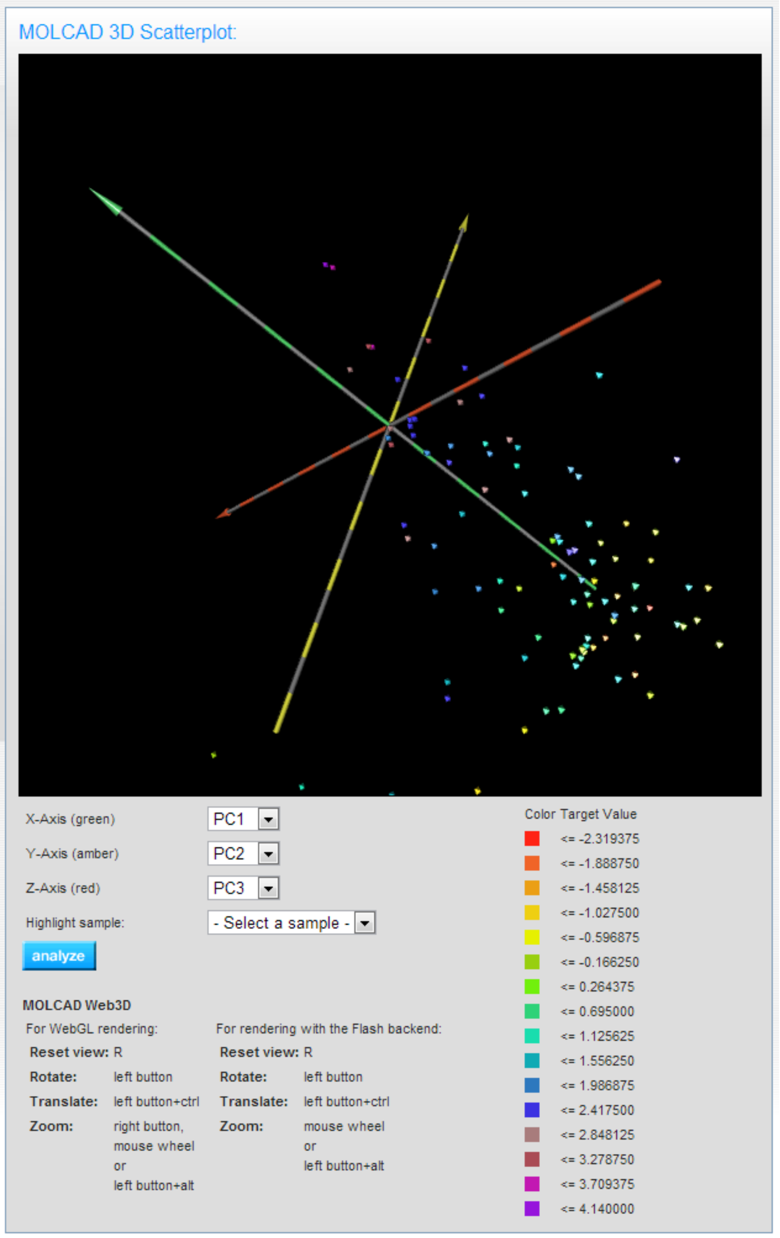
Figure 8: The first seven principal component Eigenvectors calculated for the example dataset. The coefficients of each descriptor (data column) in the analysis are given.

need to convey the information contained in the dataset?). This is shown in the SAR-caddle® PCA analysis as a plot of the eigenvalues of the principal components in decreasing order. For standardized data columns, principal components with eigenvalues larger than one contain significant information according to the widespread eigenvalue test. In the example shown, the eigenvalue plot suggests that the first ten principal components are significant. These are colored red by SAR-caddle®, as shown in Figure 7. The Eigenvector Table is shown in Figure 8. The coefficients of each of the descriptors are given, one column per principal component. The value of each principal component is obtained for each compound by multiplying the descriptor by the coefficient:

$$PC_n = \sum_{i=1}^{Ndesc} d_i a_i^n \tag{1}$$

where PC_n is the value of principal component n for the compound in question, d_i is the value of descriptor i for that compound and a_i^n is the coefficient of descriptor i in principal component n.

Because principal components analysis is a data reduction technique, plotting the values of selected principal components in either a 2D or 3D plot can reveal relationships between the data-points or clusters, and color coding the plotted points using the value of the target property can reveal relationships between it and the principal components. Figure 9 shows a SAR-caddle® interactive 3D-plot of the first three principal components for the example logP dataset color coded according to the logP value. The 3D-plot appears for all datasets that have at least three significant principal components. Otherwise, a non-interactive 2D-plot is shown. Initially, principal components 1-3 are plotted because they contain the most information. Other combinations of principal components can be selected in the three pull-down menus and a new plot requested using the analyze button. The "highlight sample" menu allows the user to select a sample to be emphasized in the plot. Once again, clicking the analyze button displays a new plot in which the requested sample (data-point) is plotted as a larger octahedron than the others. The color coding allows a fast visual estimate of how well the descriptors in the data set relate to the target property (in this case logP). If, as in this case, there is a clear gradation of the color through the plot (or if clusters with predominantly the same color are visible), the descriptors can model the target property well.


MOLCAD 3D Scatterplot:

X-Axis (green) PC1

Y-Axis (amber) PC2

Z-Axis (red) PC3

Highlight sample: - Select a sample -

Color	Target Value
🟥	<= -2.319375
🟧	<= -1.888750
🟧	<= -1.458125
🟨	<= -1.027500
🟨	<= -0.596875
🟩	<= -0.166250
🟩	<= 0.264375
🟩	<= 0.695000
🟩	<= 1.125625
🟦	<= 1.556250
🟦	<= 1.986875
🟦	<= 2.417500
🟫	<= 2.848125
🟫	<= 3.278750
🟪	<= 3.709375
🟪	<= 4.140000

MOLCAD Web3D

For WebGL rendering:		For rendering with the Flash backend:	
Reset view:	R	**Reset view:**	R
Rotate:	left button	**Rotate:**	left button
Translate:	left button+ctrl	**Translate:**	left button+ctrl
Zoom:	right button, mouse wheel or left button+alt	**Zoom:**	mouse wheel or left button+alt

Figure 9: The Molcad interactive 3D-scatter plot within SAR-caddle®. This plot can be rotated, zoomed and translated within most browsers without plugins. Internet Explorer® requires the FLASH plugin.

5.2. Shepard Interpolation

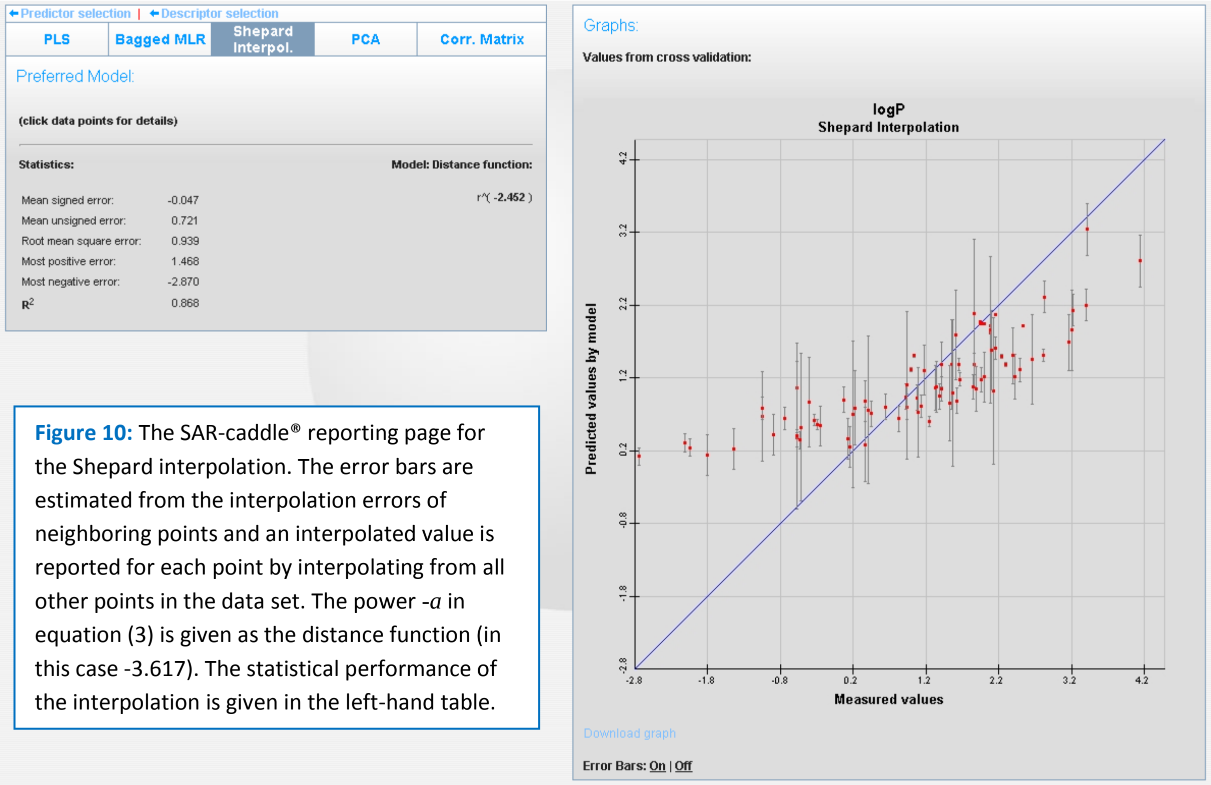
Shepard interpolation (or inverse distance weighting) is an interpolation technique that is useful for fitting data. It provides a quick view of whether the target property is related to the descriptors in principal-component space. Briefly, the target value for an unknown data point is assigned a value that is most strongly influenced by other points close to it and less strongly by points far away. The distances are calculated as the square root of sum of squares of the differences in each of the principal components between the unknown point and the neighbor:

$$R_{ij} = \sqrt{\sum_{k=1}^{N}\left(p_k^i - p_k^j\right)^2} \qquad (2)$$

where R_{ij} is the distance between data points i and j, N is the number of significant principal components used in the analysis, and p_k^i is the value of principal component k for data point i (and analogously for data point j). The interpolated value T_i for data point i (not included in the data set) is calculated as:

$$T_i = \frac{\sum_{j=1}^{N_{data}} T_j R_{ij}^{-a}}{\sum_{k=1}^{N_{data}} R_{ij}^{-a}} \qquad (3)$$

where T_j is the target value for data point j, N_{data} is the number of data points in the data set, and a is the power function for the distance dependence, which is optimized in the SAR-caddle® version of the Shepard interpolation. The results are presented as a summary table and a 2D-plot:



(click data points for details)

Statistics:

Mean signed error:	-0.047
Mean unsigned error:	0.721
Root mean square error:	0.939
Most positive error:	1.468
Most negative error:	-2.870
R^2	0.868

Model: Distance function:

r^(-2.452)

Values from cross validation:

Figure 10: The SAR-caddle® reporting page for the Shepard interpolation. The error bars are estimated from the interpolation errors of neighboring points and an interpolated value is reported for each point by interpolating from all other points in the data set. The power -*a* in equation (3) is given as the distance function (in this case -3.617). The statistical performance of the interpolation is given in the left-hand table.

The distance function provides information about the consistency of the data across the dataset. A very high negative value (-10 is the limit) means essentially that the point is assigned the value of its nearest neighbor.

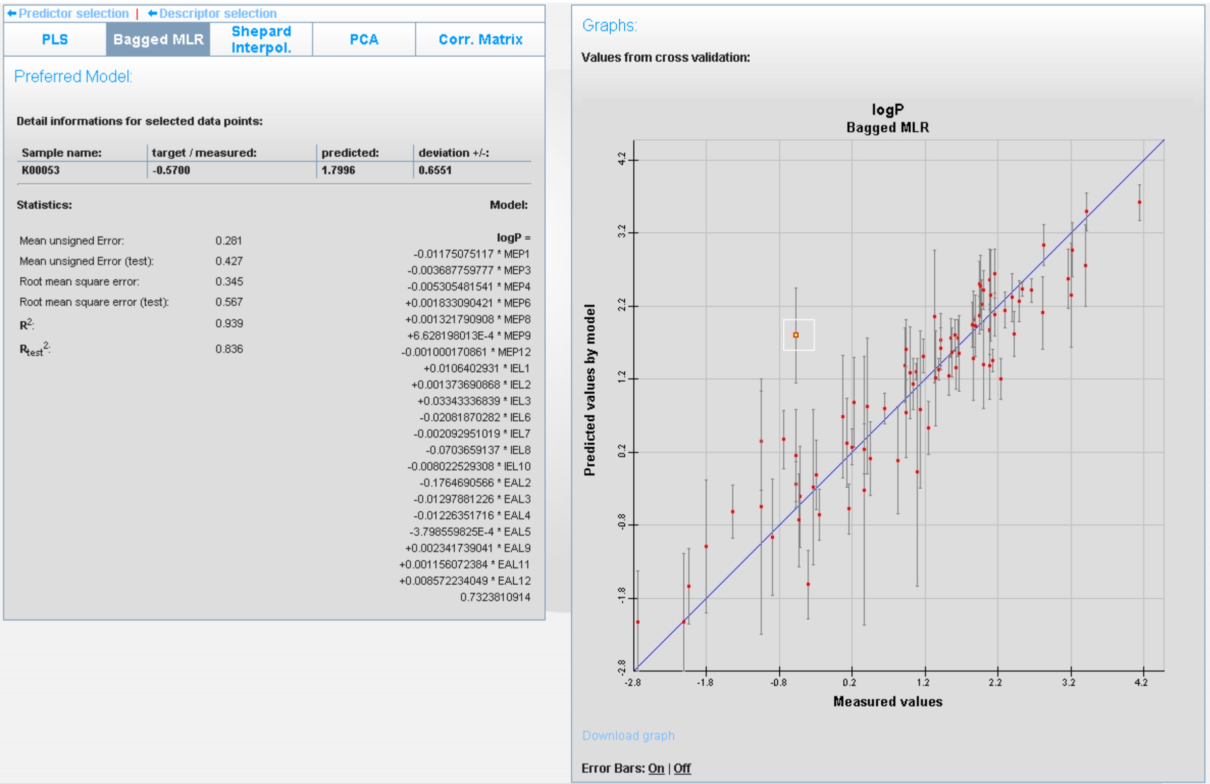
5.3. Bagged multiple linear regression (MLR)

Multiple linear regression (MLR) is a technique that models the target data as a linear combination of the descriptors:

$$T_i = c_0 + \sum_{j=1}^{N_{desc}} c_j d_j^i \qquad (4)$$

where c_0 is a constant, N_{desc} is the number of descriptors, c_j are the regression coefficients and d_j^i indicates the value of descriptor j for data point i.

The problem with such a procedure is that the higher the number of descriptors N_{desc}, the higher the possibility that the regression procedure will fit the data to random fluctuations in one or more descriptors that happen to improve the result. This results in a model that may be able to represent the training data (those used to build the model) well, but cannot predict unknown values. In order to avoid this over-training, SAR-caddle® uses two different techniques. Firstly, the F-value (the criterion used to determine whether adding another term to the regression equation is justified) is more stringent than that usually used and takes the total number of descriptors from which the algorithm can choose into account. This helps guarantee that random correlations are not included in the model.



Detail informations for selected data points:

Sample name:	target / measured:	predicted:	deviation +/-:
K00053	-0.5700	1.7996	0.6551

Statistics:

Mean unsigned Error:	0.281
Mean unsigned Error (test):	0.427
Root mean square error:	0.345
Root mean square error (test):	0.567
R^2:	0.939
R_{test}^2:	0.836

Model:

```
                          logP =
          -0.01175075117 * MEP1
         -0.003687759777 * MEP3
         -0.005305481541 * MEP4
         +0.001833090421 * MEP6
         +0.001321790908 * MEP8
         +6.628198013E-4 * MEP9
        -0.001000170861 * MEP12
            +0.0106402931 * IEL1
         +0.001373690868 * IEL2
           +0.03343336839 * IEL3
           -0.02081870282 * IEL6
          -0.002092951019 * IEL7
             -0.0703659137 * IEL8
         -0.008022529308 * IEL10
            -0.1764690566 * EAL2
           -0.01297881226 * EAL3
           -0.01226351716 * EAL4
         -3.798559825E-4 * EAL5
         +0.002341739041 * EAL9
        +0.001156072384 * EAL11
        +0.008572234049 * EAL12
                   0.7323810914
```

Values from cross validation:

Error Bars: On | Off

Figure 11: The SAR-caddle® bagged MLR output page. The error bars are derived from the distribution of the results of all the models for which the data point is in the test set. They should represent approximately ± one standard deviation.

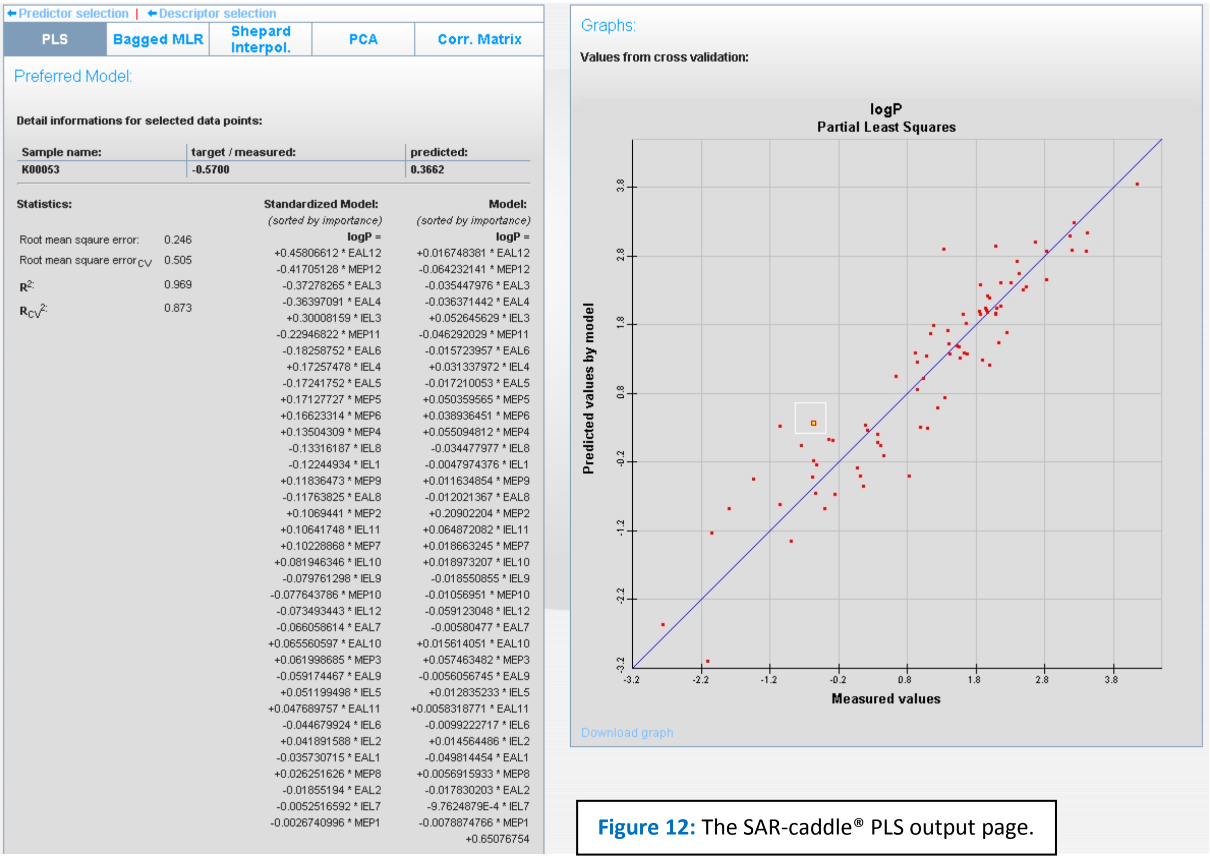
The second feature of the SAR-caddle® MLR is that it constructs many models by selecting the data points (usually about 80% of the total number in the data set) to be modeled randomly and using the remainder as the so-called test set, which is not used to build the model. This process is repeated many times and all successful models are combined to give the final model. Note that this procedure (which is known as "bagging") is stochastic and that the different test and training sets overlap. It may also happen that, for instance for small datasets some compounds never occur in a test set.

The SAR-caddle® output page for the bagged MLR, shown in Figure 11, contains a scatter plot of the test set results on the right, and the mean regression equation (the average of all the models built) together with a number of statistical performance metrics, on the left. Data points in the scatter plot may be picked with the mouse, in which case they are highlighted in white and the data for the selected point are shown at the top of the left-hand box.

5.4. Partial Least Squares Regression (PLS)

Partial least squares (PLS) regression is related to a multiple linear regression using principal components as descriptors. It describes the descriptor space as a series of orthogonal components that are analogous to principal components. The PLS algorithm used in SAR-caddle® reports the results for each number of components up to the one for which the cross-validated R^2 decreases.

Figure 12 shows the SAR-caddle® PLS results page. It is analogous to that shown above for the bagged MLR except that error bars are not available.



Preferred Model:

Detail informations for selected data points:

Sample name:	target / measured:	predicted:
K00053	-0.5700	0.3662

Statistics:

Root mean sqaure error:	0.246
Root mean square error$_{CV}$	0.505
R^2:	0.969
R_{CV}^2:	0.873

Standardized Model: (sorted by importance)	Model: (sorted by importance)
logP =	logP =
+0.45806612 * EAL12	+0.016748381 * EAL12
-0.41705128 * MEP12	-0.064232141 * MEP12
-0.37278265 * EAL3	-0.035447976 * EAL3
-0.36397091 * EAL4	-0.036371442 * EAL4
+0.30008159 * IEL3	+0.052645629 * IEL3
-0.22946822 * MEP11	-0.046292029 * MEP11
-0.18258752 * EAL6	-0.015723957 * EAL6
+0.17257478 * IEL4	+0.031337972 * IEL4
-0.17241752 * EAL5	-0.017210053 * EAL5
+0.17127727 * MEP5	+0.050359565 * MEP5
+0.16623314 * MEP6	+0.038936451 * MEP6
+0.13504309 * MEP4	+0.055094812 * MEP4
-0.13316187 * IEL8	-0.034477977 * IEL8
-0.12244934 * IEL1	-0.0047974376 * IEL1
+0.11836473 * MEP9	+0.011634854 * MEP9
-0.11763825 * EAL8	-0.012021367 * EAL8
+0.1069441 * MEP2	+0.20902204 * MEP2
+0.10641748 * IEL11	+0.064872082 * IEL11
+0.10228868 * MEP7	+0.018663245 * MEP7
+0.081946346 * IEL10	+0.018973207 * IEL10
-0.079761298 * IEL9	-0.018550855 * IEL9
-0.077643786 * MEP10	-0.01056951 * MEP10
-0.073493443 * IEL12	-0.059123048 * IEL12
-0.066058614 * EAL7	-0.00580477 * EAL7
+0.065560597 * EAL10	+0.015614051 * EAL10
+0.061998685 * MEP3	+0.057463482 * MEP3
-0.059174467 * EAL9	-0.0056056745 * EAL9
+0.051199498 * IEL5	+0.012835233 * IEL5
+0.047689757 * EAL11	+0.0058318771 * EAL11
-0.044679924 * IEL6	-0.0099222717 * IEL6
+0.041891588 * IEL2	+0.014564486 * IEL2
-0.035730715 * EAL1	-0.049814454 * EAL1
+0.026251626 * MEP8	+0.0056915933 * MEP8
-0.01855194 * EAL2	-0.017830203 * EAL2
-0.0052516592 * IEL7	-9.7624879E-4 * IEL7
-0.0026740996 * MEP1	-0.0078874766 * MEP1
	+0.65076754

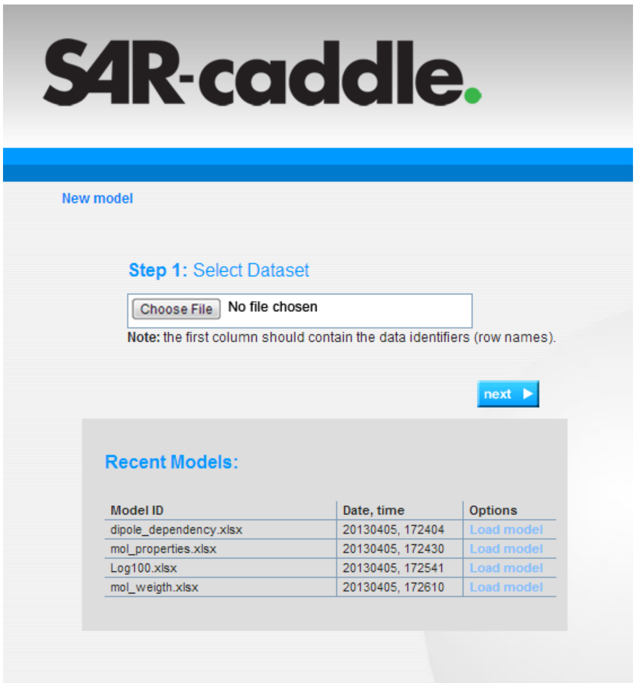
Graphs:

Values from cross validation:

Figure 12: The SAR-caddle® PLS output page.

6. Applying the models: SAR-caddle® in recall mode

When models have been made with SAR-caddle™, unknown compounds can be predicted using the recall mode. The first step is to load a model using the “new model” page, which looks like that shown in Figure 13 if models are present:



Step 1: Select Dataset

Choose File No file chosen

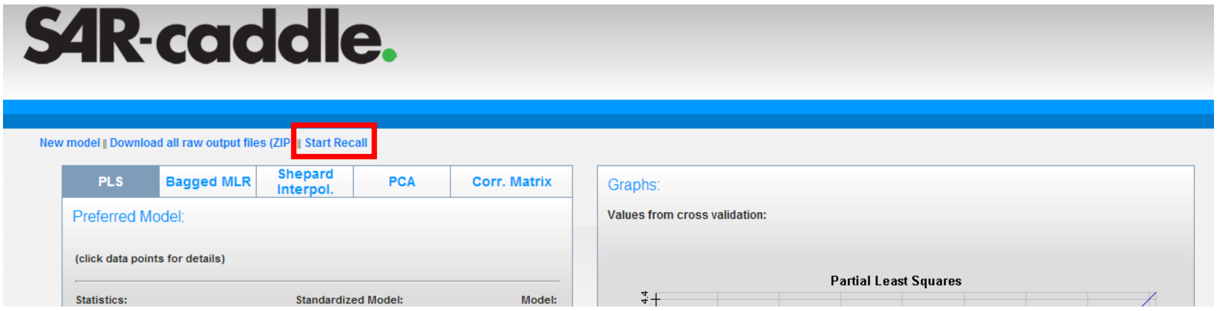
Note: the first column should contain the data identifiers (row names).

Recent Models:

Model ID	Date, time	Options
dipole_dependency.xlsx	20130405, 172404	Load model
mol_properties.xlsx	20130405, 172430	Load model
Log100.xlsx	20130405, 172541	Load model
mol_weigth.xlsx	20130405, 172610	Load model

Figure 13: The SAR-caddle® "new model" page as it appears if models have already been constructed.

Clicking on the "load model" tab for the logP_100.xlsx training dataset brings up the page with the information about the model as it appeared during training (shown in Figure 14). Clicking on the "Start Recall" tab begins the recall process.



Preferred Model:

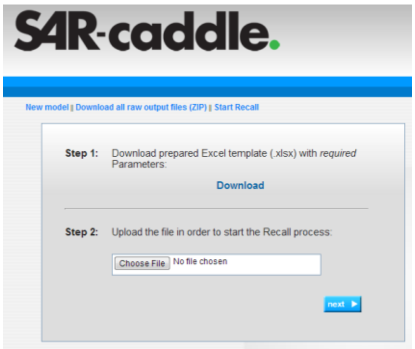
(click data points for details)

Statistics: Standardized Model: Model:

Values from cross validation:

Figure 14: The SAR-caddle® page that appears when a model is loaded. The "Start Recall" tag starts the recall (prediction of new compounds) process.

The recall process requires an input file with the descriptors (but not the predictor column) in the same order as they appeared in the training data-file. This file can either be made by downloading a template file into which the descriptors can be pasted or by loading a file that was written in advance. The relevant section of the SAR-caddle page is shown in Figure 15.


SAR-caddle.
New model | Download all raw output files (ZIP) | Start Recall
Download
next

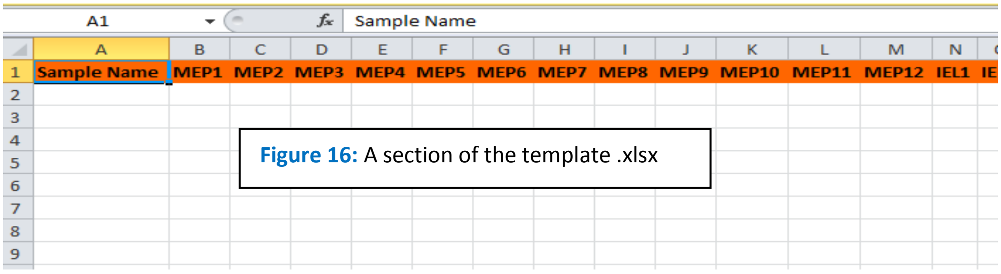
Step 1: Download prepared Excel template (.xlsx) with *required* Parameters:

Step 2: Upload the file in order to start the Recall process:

Choose File No file chosen

Figure 15: The SAR-caddle® page used to write or load an input file for the unknown compounds. The "Download" tab provides a template .xlsx file, whereas "Choose File" allows an existing file to be read in.

The file template is simply an .xlsx file in which the names of the columns have been inserted in the correct order, as shown in Figure 16.

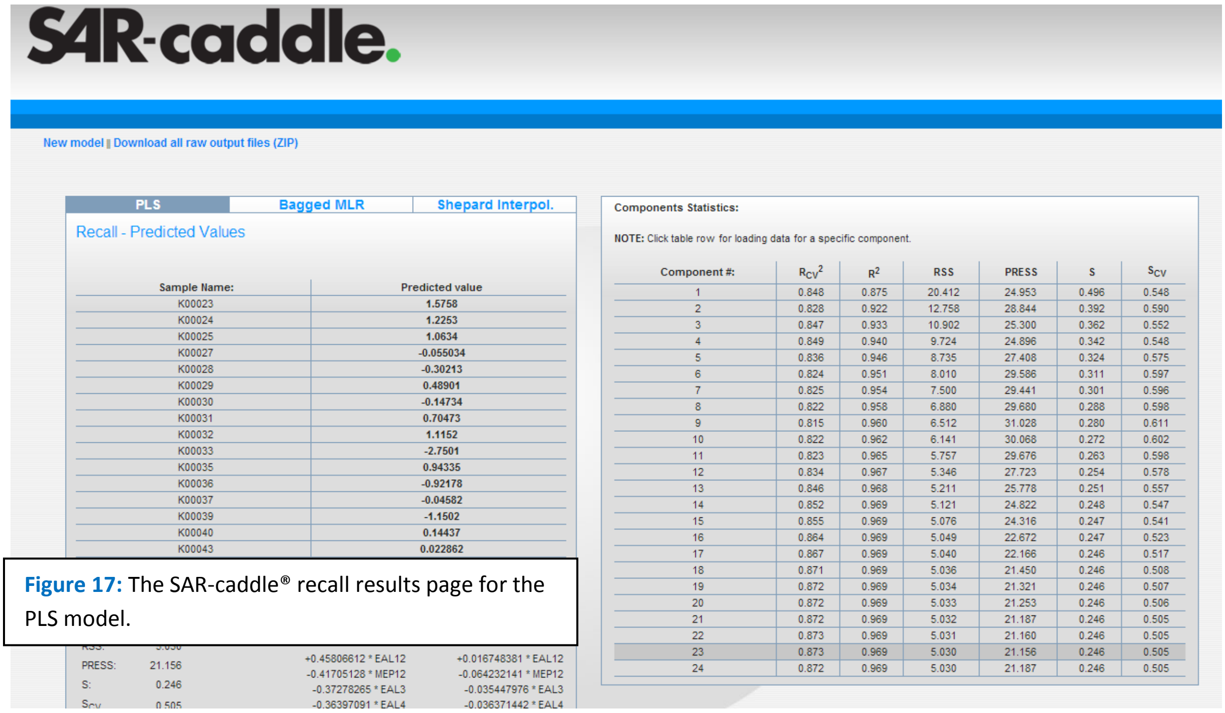


A1 fx Sample Name

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Sample Name	MEP1	MEP2	MEP3	MEP4	MEP5	MEP6	MEP7	MEP8	MEP9	MEP10	MEP11	MEP12	IEL1
2														
3														
4														
5														
6														
7														
8														
9														

Figure 16: A section of the template .xlsx

The descriptors can be entered by hand or by copy-and-paste from another file and the file can be saved. It must then be loaded using the "Choose File" tab. Choosing the file and clicking the next ▶ button applies all the available models to the new molecules. The results are shown in Figure 17.



Recall - Predicted Values

Sample Name:	Predicted value
K00023	1.5758
K00024	1.2253
K00025	1.0634
K00027	-0.055034
K00028	-0.30213
K00029	0.48901
K00030	-0.14734
K00031	0.70473
K00032	1.1152
K00033	-2.7501
K00035	0.94335
K00036	-0.92178
K00037	-0.04582
K00039	-1.1502
K00040	0.14437
K00043	0.022862

PRESS: 21.156
S: 0.246
S_{CV} 0.505

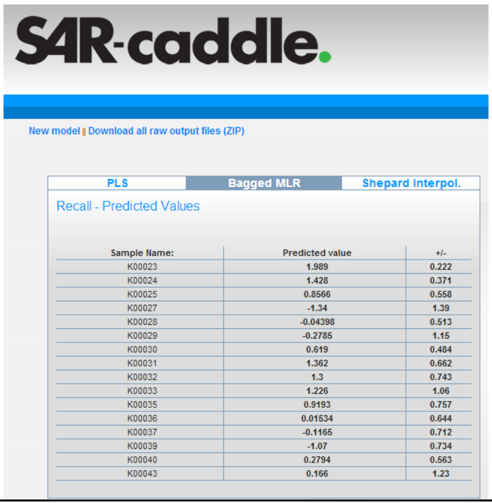
+0.45806612 * EAL12 +0.016748381 * EAL12
-0.41705128 * MEP12 -0.064232141 * MEP12
-0.37278265 * EAL3 -0.035447976 * EAL3
-0.36397091 * EAL4 -0.036371442 * EAL4

Components Statistics:

NOTE: Click table row for loading data for a specific component.

Component #:	R_{CV}^2	R^2	RSS	PRESS	S	S_{CV}
1	0.848	0.875	20.412	24.953	0.496	0.548
2	0.828	0.922	12.758	28.844	0.392	0.590
3	0.847	0.933	10.902	25.300	0.362	0.552
4	0.849	0.940	9.724	24.896	0.342	0.548
5	0.836	0.946	8.735	27.408	0.324	0.575
6	0.824	0.951	8.010	29.586	0.311	0.597
7	0.825	0.954	7.500	29.441	0.301	0.596
8	0.822	0.958	6.880	29.680	0.288	0.598
9	0.815	0.960	6.512	31.028	0.280	0.611
10	0.822	0.962	6.141	30.068	0.272	0.602
11	0.823	0.965	5.757	29.676	0.263	0.598
12	0.834	0.967	5.346	27.723	0.254	0.578
13	0.846	0.968	5.211	25.778	0.251	0.557
14	0.852	0.969	5.121	24.822	0.248	0.547
15	0.855	0.969	5.076	24.316	0.247	0.541
16	0.864	0.969	5.049	22.672	0.247	0.523
17	0.867	0.969	5.040	22.166	0.246	0.517
18	0.871	0.969	5.036	21.450	0.246	0.508
19	0.872	0.969	5.034	21.321	0.246	0.507
20	0.872	0.969	5.033	21.253	0.246	0.506
21	0.872	0.969	5.032	21.187	0.246	0.505
22	0.873	0.969	5.031	21.160	0.246	0.505
23	0.873	0.969	5.030	21.156	0.246	0.505
24	0.872	0.969	5.030	21.187	0.246	0.505

Figure 17: The SAR-caddle® recall results page for the PLS model.



Recall - Predicted Values

Sample Name:	Predicted value	+/-
K00023	1.989	0.222
K00024	1.428	0.371
K00025	0.8566	0.558
K00027	-1.34	1.39
K00028	-0.04398	0.513
K00029	-0.2785	1.15
K00030	0.619	0.484
K00031	1.362	0.662
K00032	1.3	0.743
K00033	1.226	1.06
K00035	0.9193	0.757
K00036	0.01534	0.644
K00037	-0.1165	0.712
K00039	-1.07	0.734
K00040	0.2794	0.563
K00043	0.166	1.23

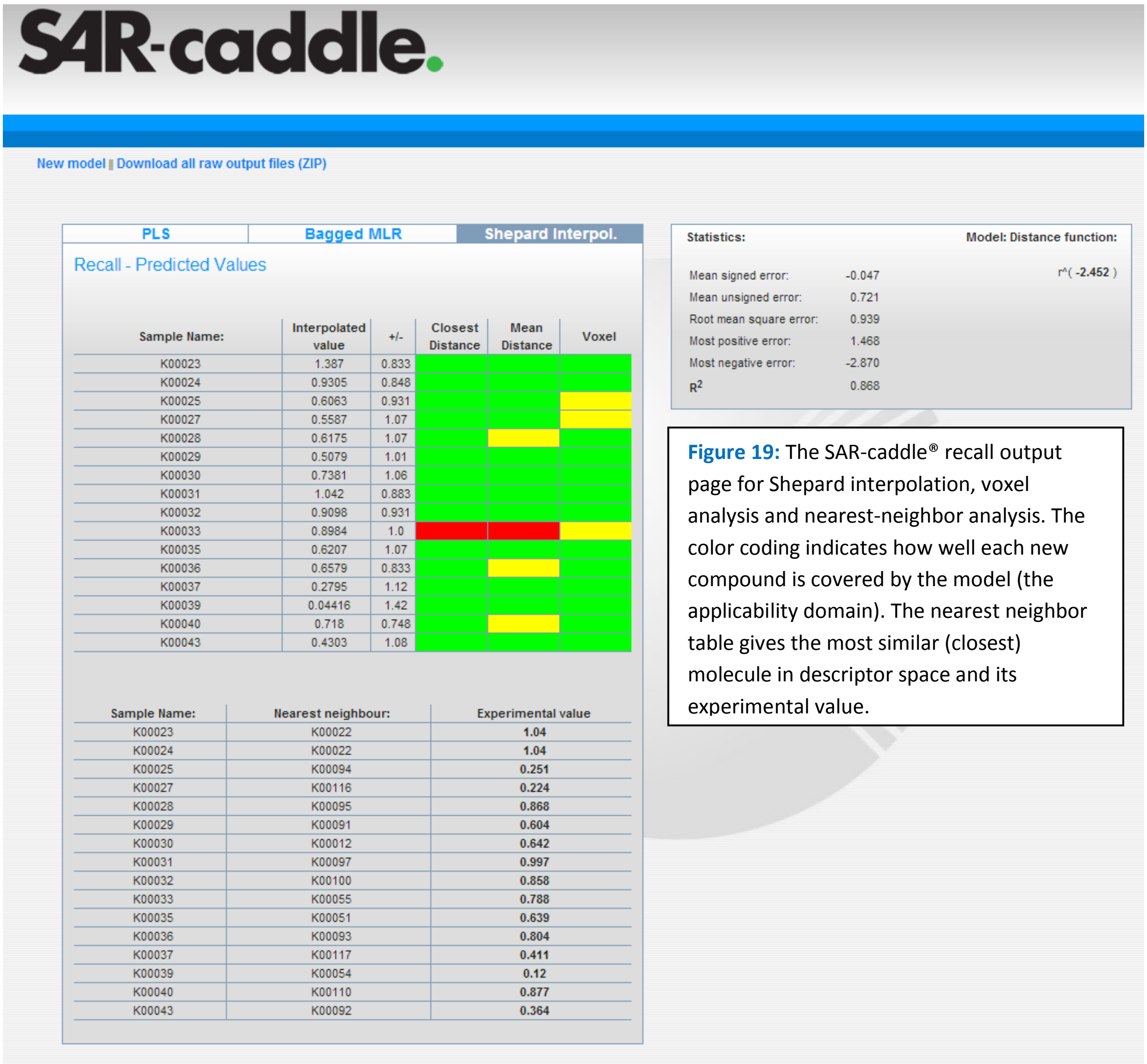
Figure 18: The SAR-caddle® bagged MLR recall page.

The first model to appear in this case is PLS. The results of the models with different numbers of components can be shown by clicking on the appropriate row in the right-hand table, exactly as for the model training. The model currently shown is marked with a darker gray background (in this case the one with six components).

The bagged MLR model results page is shown in Figure 18. The predicted value and the estimated standard deviation of the prediction is shown for each compound. These values can be compared with those predicted by the PLS model and from the Shepard interpolation model, for which the output is shown in Figure 19.

Again, the predicted value, the nearest neighbor molecule and the estimated error (standard deviation) of the prediction are shown. The probable reliability of the prediction is indicated by the three colored boxes (using the red-amber-green system), which indicate whether the voxel in which the new compound is found (“Voxel”) is well populated, the distance between the new compound and the closest one in the training set (“Closest”) and the mean distance to the training samples (“Mean”). The three measures together give an excellent indication of whether the new compound is well covered by the model. This indication also applies to the PLS and begged MLR models.

The final table shows the nearest neighbor analysis. The nearest compound (in descriptor space” and its experimental value are given.



Recall - Predicted Values

Sample Name:	Interpolated value	+/-	Closest Distance	Mean Distance	Voxel
K00023	1.387	0.833	🟩	🟩	🟩
K00024	0.9305	0.848	🟩	🟩	🟩
K00025	0.6063	0.931	🟩	🟩	🟨
K00027	0.5587	1.07	🟩	🟩	🟨
K00028	0.6175	1.07	🟩	🟨	🟩
K00029	0.5079	1.01	🟩	🟩	🟩
K00030	0.7381	1.06	🟩	🟩	🟩
K00031	1.042	0.883	🟩	🟩	🟩
K00032	0.9098	0.931	🟩	🟩	🟩
K00033	0.8984	1.0	🟥	🟥	🟨
K00035	0.6207	1.07	🟩	🟩	🟩
K00036	0.6579	0.833	🟩	🟨	🟩
K00037	0.2795	1.12	🟩	🟩	🟩
K00039	0.04416	1.42	🟩	🟩	🟩
K00040	0.718	0.748	🟩	🟨	🟩
K00043	0.4303	1.08	🟩	🟩	🟩

Sample Name:	Nearest neighbour:	Experimental value
K00023	K00022	**1.04**
K00024	K00022	**1.04**
K00025	K00094	**0.251**
K00027	K00116	**0.224**
K00028	K00095	**0.868**
K00029	K00091	**0.604**
K00030	K00012	**0.642**
K00031	K00097	**0.997**
K00032	K00100	**0.858**
K00033	K00055	**0.788**
K00035	K00051	**0.639**
K00036	K00093	**0.804**
K00037	K00117	**0.411**
K00039	K00054	**0.12**
K00040	K00110	**0.877**
K00043	K00092	**0.364**

Statistics:		Model: Distance function:
Mean signed error:	-0.047	r^(**-2.452**)
Mean unsigned error:	0.721	
Root mean square error:	0.939	
Most positive error:	1.468	
Most negative error:	-2.870	
R^2	0.868	

Figure 19: The SAR-caddle® recall output page for Shepard interpolation, voxel analysis and nearest-neighbor analysis. The color coding indicates how well each new compound is covered by the model (the applicability domain). The nearest neighbor table gives the most similar (closest) molecule in descriptor space and its experimental value.

7. Glossary Entries

Calculated Properties

The semiempirical program ParaSurf™ is able to calculate various properties from the structure of a compound. These calculated properties are particularly suited to be used as descriptors for SAR-caddle®.

Predictor

Property to be predicted by SAR-caddle®.

F-Values

For a solution of a regression task, the F-value can be calculated as

$$F = \frac{v_2 R^2}{v_1 \left(1 - R^2\right)}$$

C. Kramer, C. S. Tautermann, C. Kramer, D. J. Livingstone, D. W. Salt, D. C. Whitley, B. Beck and T. Clark, J. Chem. Inf. Mod. **49**, 28-34, 2009. doi: 10.1021/ci800318q

Test set

A test set is a set of data used to provide an independent estimate of the predictive ability of a model. These data fit within the applicability domain of the model, but have not been used to train it.

Bagging

Each model consists of 100* independent multiple linear regression models that were built based on randomly chosen 75%* fractions of the overall data set. The remaining 25%* of the data set are used as a test set. On average every compound therefore occurs 25 times in the test set.

*default values

Eigenvalue test

All eigenvalues for the correlation matrix are computed and all factors with eigenvalues under 1.0 are dropped. All factors with eigenvalues greater than one are included in the model.