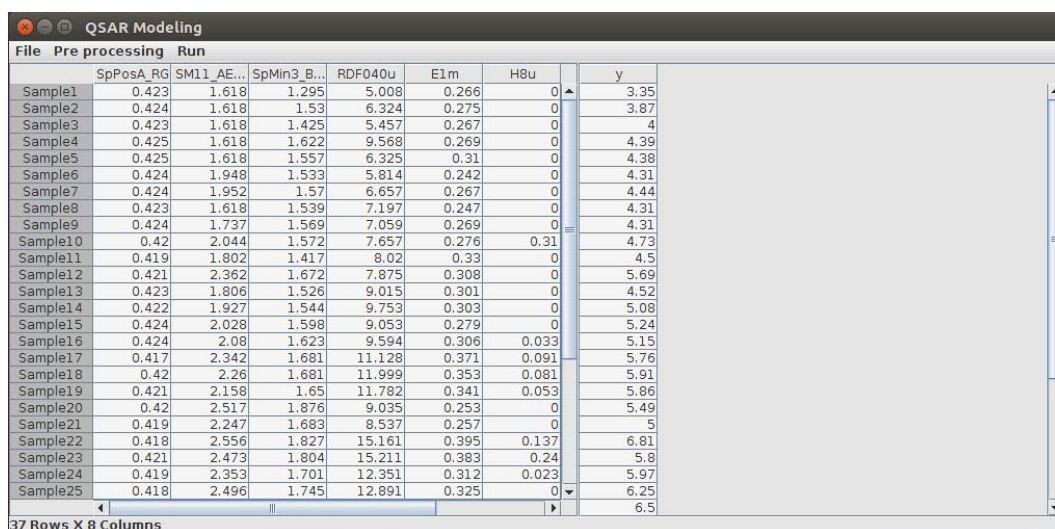


## Supplementary Information

***QSAR modeling*: a new open source computational package to generate and validate QSAR models**

**João Paulo A. Martins e Márcia M. C. Ferreira\*** [marcia@iqm.unicamp.br](mailto:marcia@iqm.unicamp.br)

Instituto de Química, Universidade Estadual de Campinas, CP 6154, 13084-971 Campinas – SP, Brasil

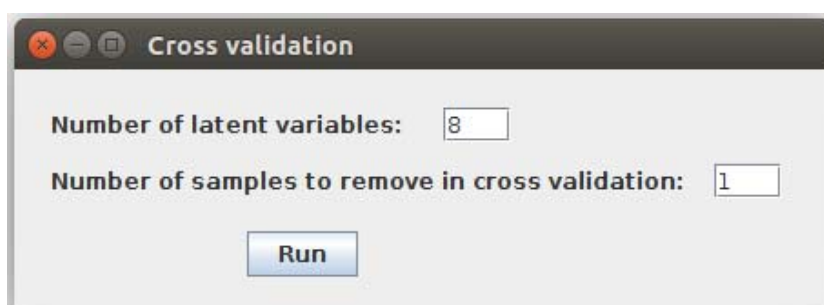


The screenshot shows the main window of the QSAR Modeling software. The window has a menu bar with 'File', 'Pre processing', and 'Run'. Below the menu bar is a table with 25 rows (labeled Sample1 to Sample25) and 8 columns. The columns are: SpPosA\_RG, SM11\_AE..., SpMin3\_B..., RDF040u, Elm, H8u, and y. The data is as follows:

	SpPosA_RG	SM11_AE...	SpMin3_B...	RDF040u	Elm	H8u	y
Sample1	0.423	1.618	1.295	5.008	0.266	0	3.35
Sample2	0.424	1.618	1.53	6.324	0.275	0	3.87
Sample3	0.423	1.618	1.425	5.457	0.267	0	4
Sample4	0.425	1.618	1.622	9.568	0.269	0	4.39
Sample5	0.425	1.618	1.557	6.325	0.31	0	4.38
Sample6	0.424	1.948	1.533	5.814	0.242	0	4.31
Sample7	0.424	1.952	1.57	6.657	0.267	0	4.44
Sample8	0.423	1.618	1.539	7.197	0.247	0	4.31
Sample9	0.424	1.737	1.569	7.059	0.269	0	4.31
Sample10	0.42	2.044	1.572	7.657	0.276	0.31	4.73
Sample11	0.419	1.802	1.417	8.02	0.33	0	4.5
Sample12	0.421	2.362	1.672	7.875	0.308	0	5.69
Sample13	0.423	1.806	1.526	9.015	0.301	0	4.52
Sample14	0.422	1.927	1.544	9.753	0.303	0	5.08
Sample15	0.424	2.028	1.598	9.053	0.279	0	5.24
Sample16	0.424	2.08	1.623	9.594	0.306	0.033	5.15
Sample17	0.417	2.342	1.681	11.128	0.371	0.091	5.76
Sample18	0.42	2.26	1.681	11.999	0.353	0.081	5.91
Sample19	0.421	2.158	1.65	11.782	0.341	0.053	5.86
Sample20	0.42	2.517	1.876	9.035	0.253	0	5.49
Sample21	0.419	2.247	1.683	8.537	0.257	0	5
Sample22	0.418	2.556	1.827	15.161	0.395	0.137	6.81
Sample23	0.421	2.473	1.804	15.211	0.383	0.24	5.8
Sample24	0.419	2.353	1.701	12.351	0.312	0.023	5.97
Sample25	0.418	2.496	1.745	12.891	0.325	0	6.25

At the bottom of the window, it says '37 Rows X 8 Columns'.

Figure 1S: The main screen of *QSAR modeling* program



The screenshot shows a 'Cross validation' dialog box. It has two input fields: 'Number of latent variables:' with the value '8' and 'Number of samples to remove in cross validation:' with the value '1'. Below these fields is a 'Run' button.

Figure 2S. Window from *QSAR Modeling* where the user chooses the maximum number of latent variables and the number of samples to be removed during cross validation.

The image shows a software window titled "OPS" with standard window controls (close, minimize, maximize). The window contains several input fields and two groups of radio buttons. The input fields are for "Number of latent variables for OPS:", "Number of latent variables for the model:", "Number of samples to remove in cross validation:", "Window:", "Increment:", and "Percentage of variables:". Below these are two groups of radio buttons. The first group, labeled "Vector", has three options: "Correlogram" (selected), "Regression vector", and "Product". The second group, labeled "Criterion to classify the m...", has four options: "RMSECV" (selected), "Rcv", "Q2", and "Spres". At the bottom center of the window is a "Run" button.

OPS

Number of latent variables for OPS:

Number of latent variables for the model:

Number of samples to remove in cross validation:

Window:  Increment:

Percentage of variables:

**Vector**

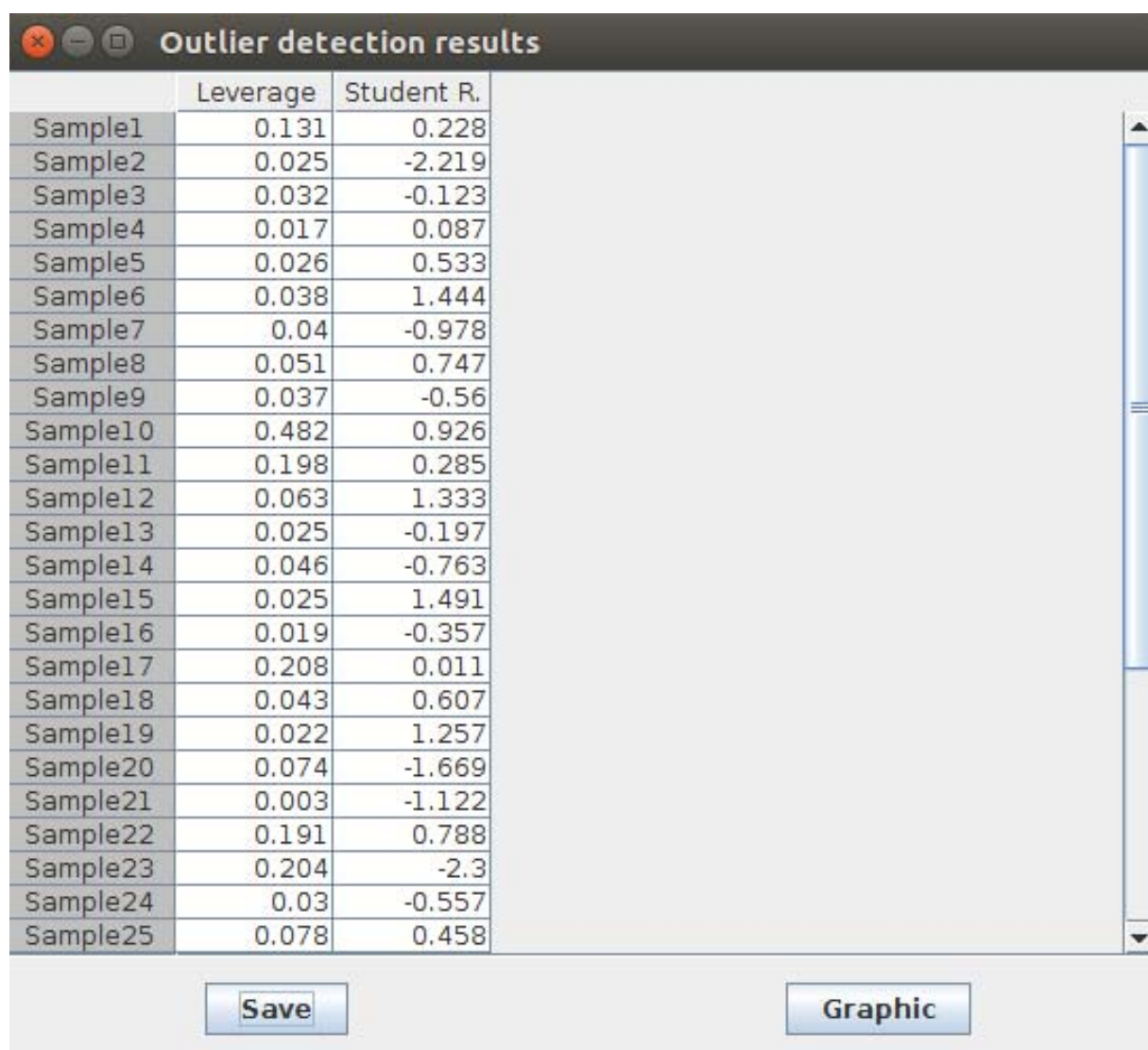
- ☒ Correlogram
- ☐ Regression vector
- ☐ Product

**Criterion to classify the m...**

- ☒ RMSECV
- ☐ Rcv
- ☐ Q2
- ☐ Spres

Run

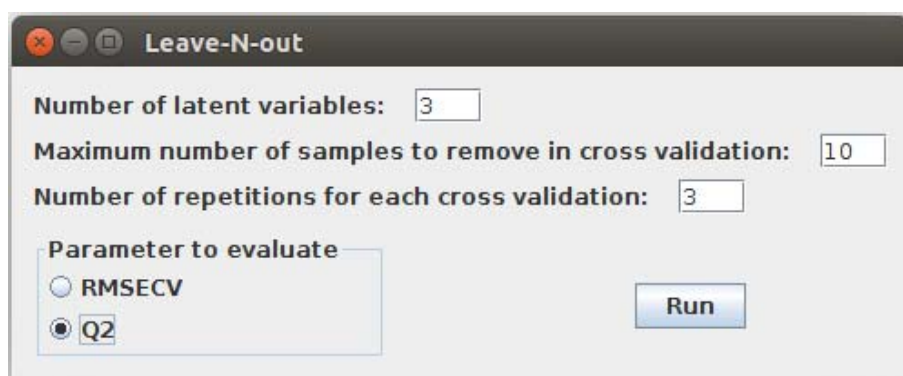
Figure 3S: *QSAR modeling* window where the user can choose the options to run the OPS algorithm



	Leverage	Student R.
Sample1	0.131	0.228
Sample2	0.025	-2.219
Sample3	0.032	-0.123
Sample4	0.017	0.087
Sample5	0.026	0.533
Sample6	0.038	1.444
Sample7	0.04	-0.978
Sample8	0.051	0.747
Sample9	0.037	-0.56
Sample10	0.482	0.926
Sample11	0.198	0.285
Sample12	0.063	1.333
Sample13	0.025	-0.197
Sample14	0.046	-0.763
Sample15	0.025	1.491
Sample16	0.019	-0.357
Sample17	0.208	0.011
Sample18	0.043	0.607
Sample19	0.022	1.257
Sample20	0.074	-1.669
Sample21	0.003	-1.122
Sample22	0.191	0.788
Sample23	0.204	-2.3
Sample24	0.03	-0.557
Sample25	0.078	0.458

Save Graphic

Figure 4S. Results from outlier detection including Leverage and Studentized residuals values for compounds from the training set.



A dialog box titled "Leave-N-out" with a standard window header (close, minimize, maximize buttons). It contains three input fields: "Number of latent variables:" with the value 3, "Maximum number of samples to remove in cross validation:" with the value 10, and "Number of repetitions for each cross validation:" with the value 3. Below these is a section titled "Parameter to evaluate" with two radio buttons: "RMSECV" (unselected) and "Q2" (selected). A "Run" button is located to the right of the radio buttons.

Number of latent variables: 3

Maximum number of samples to remove in cross validation: 10

Number of repetitions for each cross validation: 3

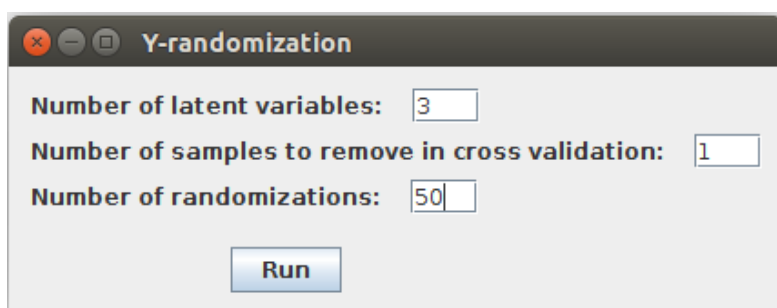
Parameter to evaluate

☐ RMSECV

☒ Q2

Run

Figure 5S: Leave- $N$ -out validation procedure to assess the robustness of a model using *QSAR modeling*.



A dialog box titled "Y-randomization" with a standard window header (close, minimize, maximize buttons). It contains three input fields: "Number of latent variables:" with the value 3, "Number of samples to remove in cross validation:" with the value 1, and "Number of randomizations:" with the value 50. A "Run" button is located at the bottom center.

Number of latent variables: 3

Number of samples to remove in cross validation: 1

Number of randomizations: 50

Run

Figure 6S.  $y$ -randomization validation procedure to verify chance correlation of a model using *QSAR modeling*.