# Supplementary Data

## Critical comparative analysis, validation and interpretation of SVM and PLS regression models in a QSAR study on HIV-1 protease inhibitors

Noslen Hernández[a], Rudolf Kiralj[b], Márcia M. C. Ferreira[b]*, Isneri Talavera[a]

[a]Advanced Technologies Application Center, Havana, 12200, Cuba

[b]Instituto de Química, Universidade Estadual de Campinas, Campinas, SP 13083-970, Brazil

## CONTENTS

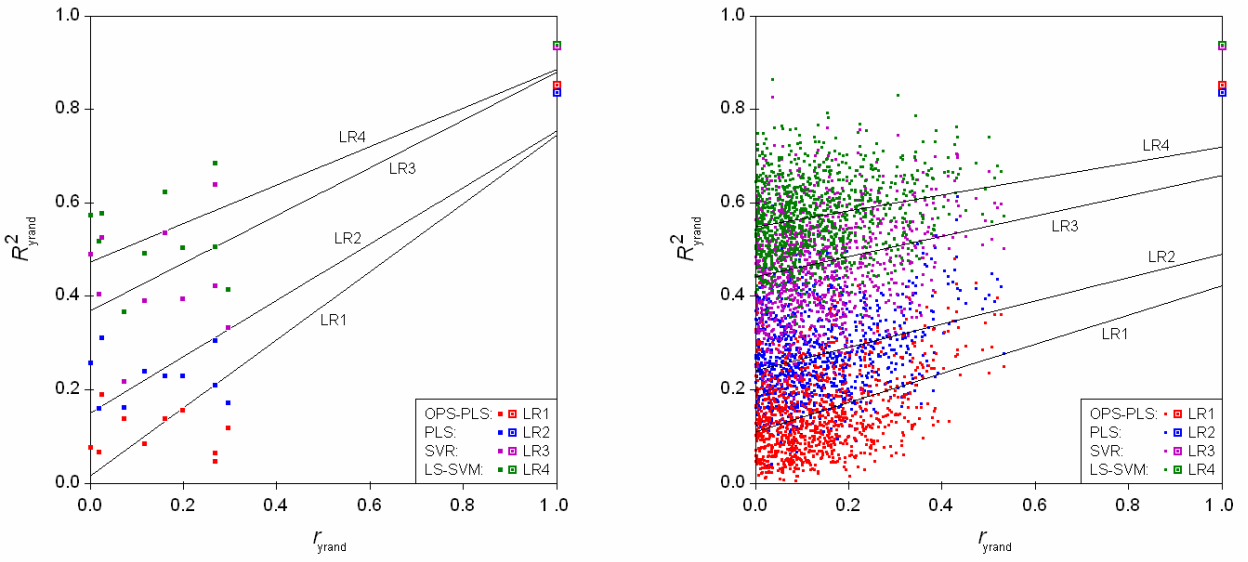


Figure S1. The $R^2_{yrand}$ against $r_{yrand}$ plots for 10 (left) and 1000 (right) randomizations of the four QSAR models, with the corresponding linear regression (LR) lines. The proposed QSAR models are situated at the right upper corner and are marked by larger symbols.

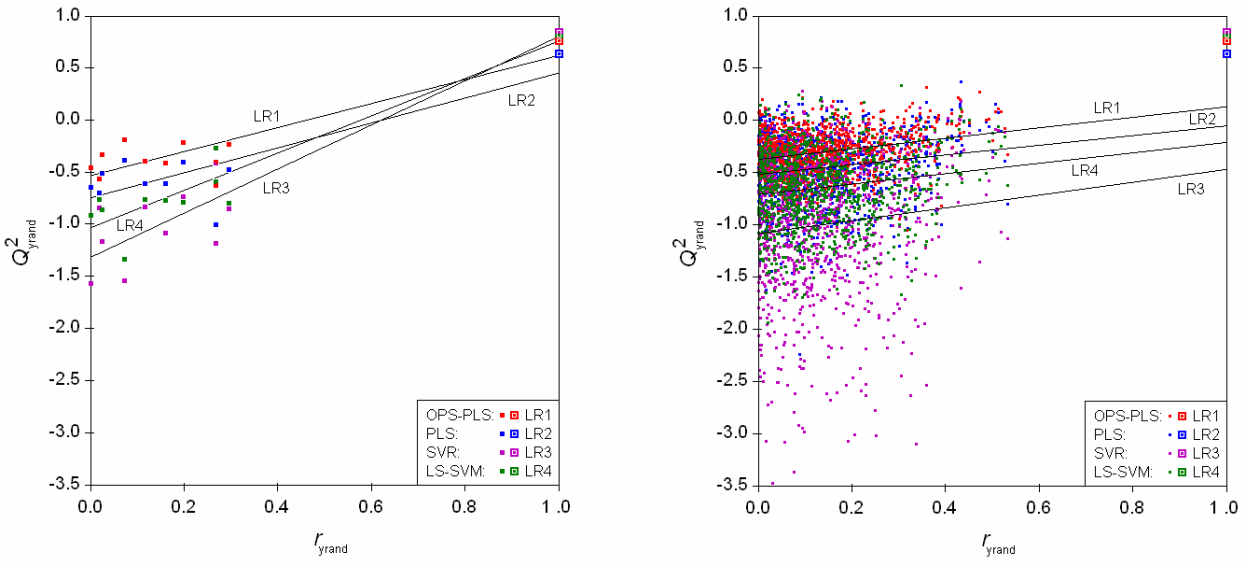


Figure S2. The $Q^2_{yrand}$ against $r_{yrand}$ plots for 10 (left) and 1000 (right) randomizations of the four QSAR models, with the corresponding linear regression (LR) lines. The proposed QSAR models are situated at the right upper corner and are marked by larger symbols.

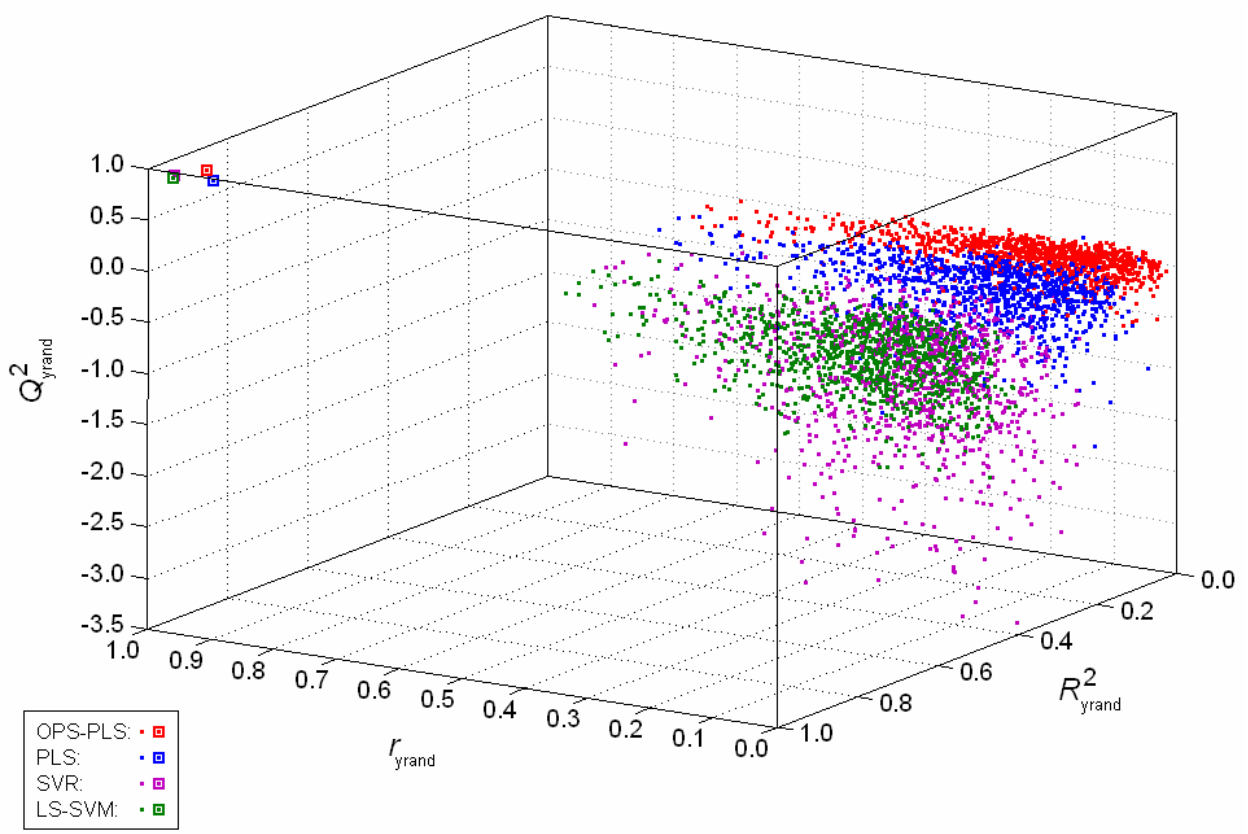


Figure S3. The 3D plot $r_{yrand}$-$R^2_{yrand}$-$Q^2_{yrand}$ for 1000 randomizations of the four QSAR models. The proposed QSAR models are situated at the left upper corner and are marked by larger symbols.

Table S1. Linear regression (LR) equations from **y**-randomization validations of the four QSAR models.

| Plot | QSAR Model | $M$[a] | LR Equation[b] | $\Delta$[c] |
|---|---|---|---|---|
| $Q^2_{yrand}$ - $R^2_{yrand}$ | PLS | 10 | LR2: $Q^2_{yrand}$ = -0.984(135) + 1.778(400) $R^2_{yrand}$ | 0.01 |
| | | 1000 | LR2: $Q^2_{yrand}$ = -0.986 (25) + 1.935(86) $R^2_{yrand}$ | 0.38 |
| | OPS-PLS | 10 | LR1: $Q^2_{yrand}$ = -0.554(41) + 1.564(147) $R^2_{yrand}$ | 0.48 |
| | | 1000 | LR1: $Q^2_{yrand}$ = -0.574(8) + 1.763(46) $R^2_{yrand}$ | 1.29 |
| | SVR | 10 | LR3: $Q^2_{yrand}$ = -2.207(357) + 2.821(696) $R^2_{yrand}$ | 0.32 |
| | | 1000 | LR3: $Q^2_{yrand}$ = -2.089(82) + 2.311(170) $R^2_{yrand}$ | 0.71 |
| | LS-SVM | 10 | LR4: $Q^2_{yrand}$ = -2.459(289) + 3.229(496) $R^2_{yrand}$ | 1.34 |
| | | 1000 | LR4: $Q^2_{yrand}$ = -2.062(64) + 2.484(111) $R^2_{yrand}$ | 1.47 |
| $Q^2_{yrand}$ - $r_{yrand}$ | PLS | 10 | LR2: $Q^2_{yrand}$ = -0.746(97) + 1.196(279) $r_{yrand}$ | 0.43 |
| | | 1000 | LR2: $Q^2_{yrand}$ = -0.518(15) + 0.464(87) $r_{yrand}$ | 2.50 |
| | OPS-PLS | 10 | LR1: $Q^2_{yrand}$ = -0.534(76) + 1.153(220) $r_{yrand}$ | 2.09 |
| | | 1000 | LR1: $Q^2_{yrand}$ = -0.374(10) + 0.503(57) $r_{yrand}$ | 2.86 |
| | SVR | 10 | LR3: $Q^2_{yrand}$ = -1.318(119) + 2.118(344) $r_{yrand}$ | 1.90 |
| | | 1000 | LR3: $Q^2_{yrand}$ = -1.084(31) + 0.613(173) $r_{yrand}$ | 3.91 |
| | LS-SVM | 10* | LR4: $Q^2_{yrand}$ = -1.033(88) + 1.790(255) $r_{yrand}$ | 3.58 |
| | | 1000* | LR4: $Q^2_{yrand}$ = -0.711(18) + 0.498(100) $r_{yrand}$ | 4.72 |
| $R^2_{yrand}$ - $r_{yrand}$ | PLS | 10 | LR2: $R^2_{yrand}$ = 0.150(36) + 0.603(105) $r_{yrand}$ | 2.50 |
| | | 1000 | LR2: $R^2_{yrand}$ = 0.241(5) + 0.249(25) $r_{yrand}$ | 3.28 |
| | OPS-PLS | 10 | LR1: $R^2_{yrand}$ = 0.015(43) + 0.729(124) $r_{yrand}$ | 2.20 |
| | | 1000 | LR1: $R^2_{yrand}$ = 0.110(4) + 0.312(24) $r_{yrand}$ | 3.30 |
| | SVR | 10 | LR3: $R^2_{yrand}$ = 0.368(50) + 0.510(145) $r_{yrand}$ | 1.45 |
| | | 1000 | LR3: $R^2_{yrand}$ = 0.441(5) + 0.217(29) $r_{yrand}$ | 1.98 |
| | LS-SVM | 10 | LR4: $R^2_{yrand}$ = 0.473(42) + 0.412(120) $r_{yrand}$ | 1.78 |
| | | 1000 | LR4: $R^2_{yrand}$ = 0.548(4) + 0.171(23) $r_{yrand}$ | 1.97 |

[a]Number of **y**-randomization runs.

[b]Statistical errors on regression coefficients are given in brackets for the last three digits.

[c]Differences between regression equations for 10 and 1000 randomizations in terms of regression coefficients are calculated as follows: $\Delta = [p_1 - p_2] / [\sigma(p_1)^2 + \sigma(p_2)^2]^{1/2}$, where $p_1$ and $p_2$ are the values of a particular regression coefficient from the two equations, and $\sigma(p_1)$ and $\sigma(p_2)$ are the respective errors. For each pair of equations, the top and bottom values refer to the free and linear coefficients, respectively.

*Pair of equations with at least one extremely significant difference in regression coefficients, *i.e.*, $\Delta > 3.89$ what corresponds to the confidence level <0.0001, assuming that the differences are normally distributed.

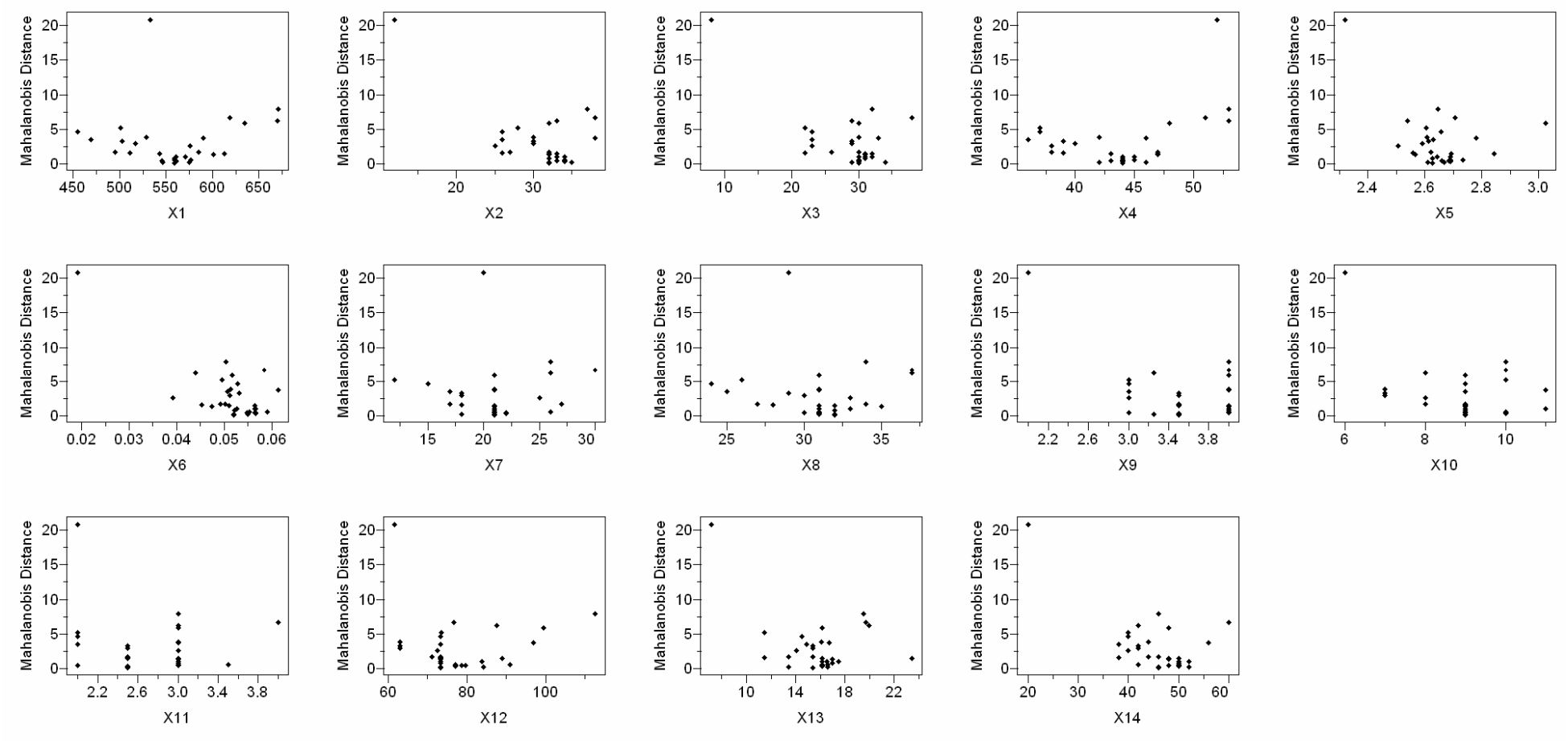


Figure S4. The descriptor against Mahalanobis distance scatterplots showing different types of relationships, what mainly corresponds to four HCA clusters. Descriptors $X_1$, $X_4$, $X_7$ and $X_8$ are well correlated to Mahalanobis distance over the whole descriptor range (with exception of one sample), meaning that they do not bring new information because of which they would be included in variable selection carried out by the OPS-PLS procedure.