Electronic Supplementary Material (ESI) for Environmental Science: Processes & Impacts.

Prediction of Acute Toxicity of Emerging Contaminants on the Water Flea *Daphnia magna* by Ant Colony Optimization - Support Vector Machine QSTR models

Reza Aalizadeh †, Peter C. von der Ohe ‡ and Nikolaos S. Thomaidis *, †

† Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771 Athens, Greece

‡ German Environment Agency, 06844 Dessau-Roßlau, Germany

SI1. Supporting information for Validation criteria

Internal evaluation parameters were calculated to compare the proposed models and obtain the best model. R^2 (squared correlation coefficient) and Q^2_{LOO} are used for that purpose that are calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I}(y_i - \bar{y})^2} \qquad (Eq.1)$$

Where y_i are experimental properties, $\bar{y}$ is the average of all y_i, and $\hat{y}_i$ are the predicted values. However, for calculating the Q^2_{LOO} values, firstly one of compounds in the dataset is being excluded and its property is being calculated by the (M-1×N) model (M is the number of compounds and N is the number of descriptors). This process continues until every compound in the dataset has been excluded once and then the correlation coefficients for newly predicted values and experimental properties are being calculated. Furthermore, root mean square errors (RMSE), variation inflation factors (VIF), and Lin's concordance correlation coefficient (CCC) [1] were assessed for better comparison. CCC values inspect the degree to which the pairs of data points fall onto the 45° (1:1) line through the origin. CCC values were calculated as below:

$$CCC = \frac{2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{x} - y)^2} \qquad (Eq.2)$$

where, x and y are abscissa and ordinate value of the graph plotting the experimental values versus the ones predicted by the mode1, respectively. n is number of compounds in the dataset, $\bar{x}$ and $\bar{y}$ are the averages of x and y, respectively. The CCC value is also a robust evaluator for the external evaluation of a model.

Several other techniques were used as well to evaluate the external capability of the proposed models. In addition to the statistical parameters introduced above (R^2, RMSE, and CCC),

modified r^2 (r^2_m) [2], Q_{F1}^2 [3], Q_{F2}^2 [3], Q_{F3}^2 [4] [5] are acceptable validation criteria suggested by Golbraikh and Tropsha [6]. Modified r^2 values can be calculated for both training and test set as follows:

$$r_m^2 = R^2\left(1 - \sqrt{\left(R^2 - R_0^2\right)}\right) \qquad (Eq.4)$$

where, R_0^2 is the correlation coefficient with intercept of zero. Moreover, Q_{F1}^2, Q_{F2}^2 and Q_{F3}^2, as OECD principles, are being calculated as follows:

$$Q_{F1}^2 = 1 - \frac{\sum_{i=1}^{n(ext)} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n(ext)} (y_i - \bar{y}_{train})^2} \qquad (Eq.5)$$

$$Q_{F2}^2 = 1 - \frac{\sum_{i=1}^{n(ext)} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n(ext)} (y_i - \bar{y}_{ext})^2} \qquad (Eq.6)$$

$$Q_{F3}^2 = 1 - \frac{\left[\sum_{i=1}^{n(ext)} (\hat{y}_i - y_i)^2\right]/n_{ext}}{\left[\sum_{i=1}^{n(train)} (y_i - \bar{y}_{train})^2\right]/n_{train}} \qquad (Eq.7)$$

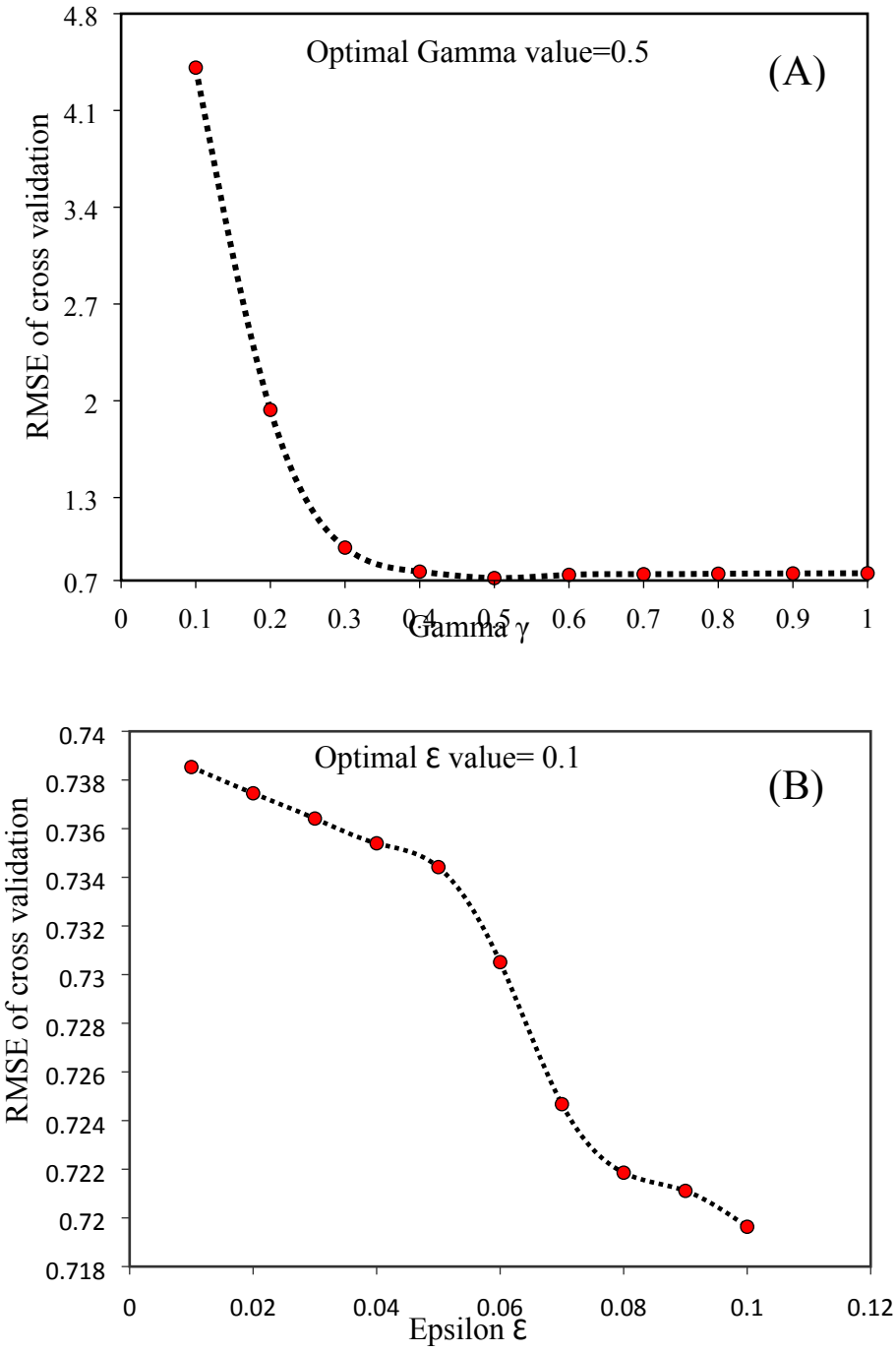
where $n_{(ext)}$is the number of compounds in test set, y_i and $\hat{y}_i$ are the observed and predicted values of the measured property in the test set. $\bar{y}_{train}$and $\bar{y}_{ext}$are the average values of the observed and predicted properties in the training and test set, respectively.

Based on Golbraikh and Tropsha's models acceptance criteria [6, 7], the model criteria proposed by Erikson et al. [8], and all above validation methods, a model can be regarded as acceptable for prediction purposes if it meets the following conditions:

1. $Q^2_{LOO} > 0.5$
2. $R^2 > 0.6$
3. $\left(R^2 - R_0^2\right)/R^2 < 0.1 \quad or \left(R^2 - R_0^{'2}\right)/R^2 < 0.1$ ($R_0^{'2}$ value is calculated by change of axes, k and $k^{'}$ are the slops and in case of axes changes, namely)
4. $0.85 \le k \le 1.15 \quad or \quad 0.85 \le k^{'} \le 1.15$
5. $r_m^2 > 0.5$
6. $CCC > 0.85$
7. $\left| R^2_{Training} - R^2_{Test} \right| < 0.3$
8. Q^2_{F1} & Q^2_{F2} & $Q^2_{F3} > 0.6$

SI2. Results and discussion

To derive a robust ACO-SVM model, the internal parameter of SVM was optimized using RMSE of Q^2_{LOO}. The lowest RMSE of Q^2_{LOO} was observed at C=3, ε=0.1 and γ=0.5 (**Fig. S1**)



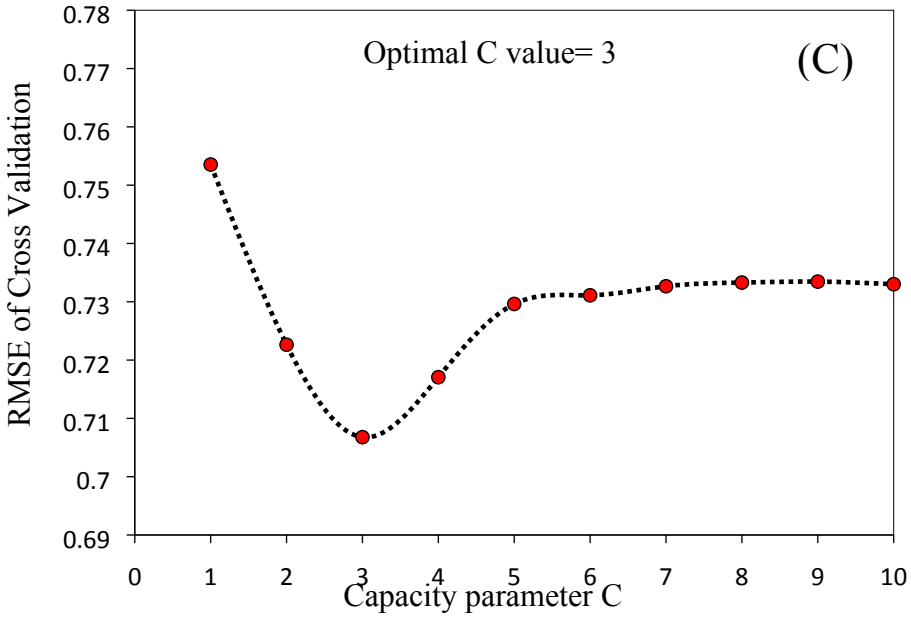
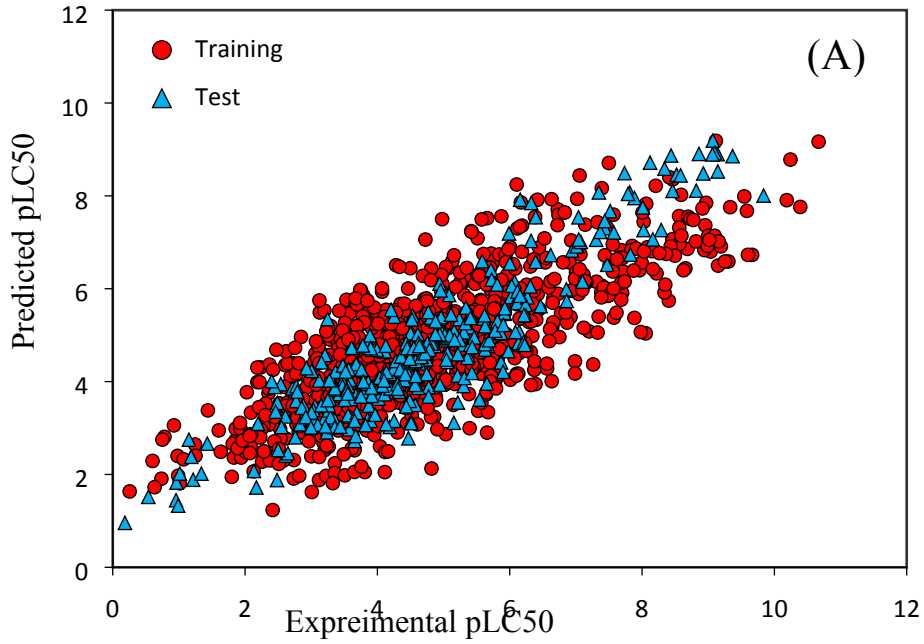


Fig. S1 Optimization of γ, ε and C parameters in SVM

The correlations between the predicted and the experimental toxicity for ACO-MLR and optimized ACO-SVM are shown in **Fig S2**.



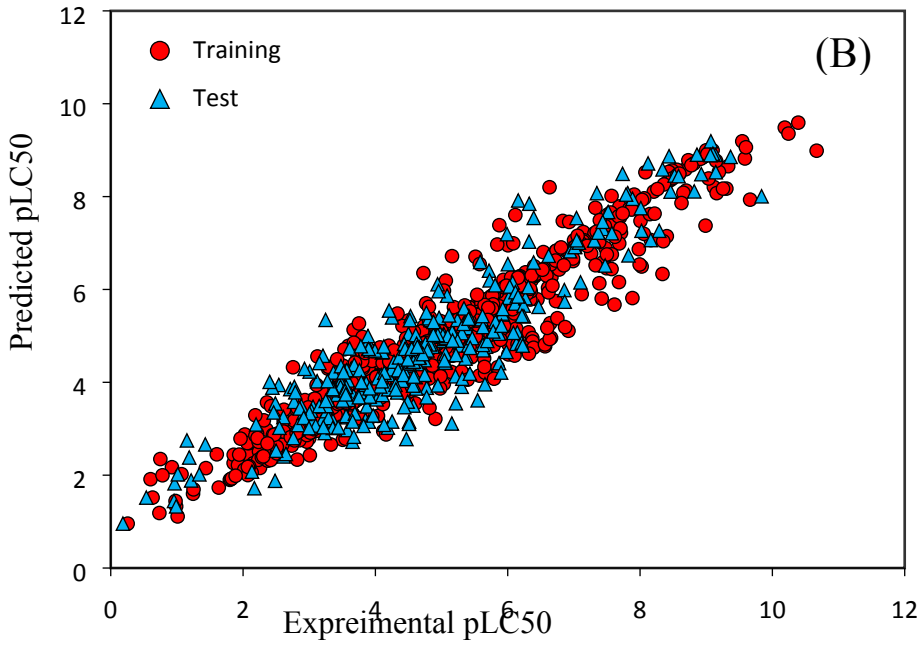


Fig. S2 Correlation between experimental and predicted Toxicity: A) ACO-MLR and B) ACO-SVM

Reference

1. Lin, L.I., *A Concordance Correlation Coefficient to Evaluate Reproducibility.* Biometrics, 1989. **45**(1): p. 255-68.
2. Roy, K., et al., *Some case studies on application of "rm2" metrics for judging quality of quantitative structure–activity relationship predictions: Emphasis on scaling of response data.* Journal of Computational Chemistry, 2013. **34**(12): p. 1071-1082.
3. Shi, L.M., et al., *QSAR Models Using a Large Diverse Set of Estrogens.* Journal of Chemical Information and Computer Sciences, 2000. **41**(1): p. 186-195.
4. Consonni, V., D. Ballabio, and R. Todeschini, *Comments on the Definition of the Q2 Parameter for QSAR Validation.* Journal of Chemical Information and Modeling, 2009. **49**(7): p. 1669-1678.
5. Singh, K.P. and S. Gupta, *Nano-QSAR modeling for predicting biological activity of diverse nanomaterials.* RSC Advances, 2014. **4**(26): p. 13215-13230.
6. Golbraikh, A. and A. Tropsha, *Beware of q2!* Journal of Molecular Graphics and Modelling, 2002. **20**(4): p. 269-276.
7. Tropsha, A., *Best Practices for QSAR Model Development, Validation, and Exploitation.* Molecular Informatics, 2010. **29**(6-7): p. 476-488.
8. Eriksson, L., et al., *Methods for Reliability and Uncertainty Assessment and for Applicability Evaluations of Classification- and Regression-Based QSARs.* Environmental Health Perspectives 2003. **111**(10): p. 1361-75.