QSRR models to support suspect high resolution mass spectrometric screening of emerging contaminants in environmental samples

Reza Aalizadeh, Nikolaos S. Thomaidis*, Anna A. Bletsou and Pablo Gago Ferrero

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

* Corresponding author:

Tel: +30 210 7274317

Fax: +30 210 7274750

E-mail: ntho@chem.uoa.gr

Submitted to: Journal of Chemical Information and Modeling

ABSTRACT

Over the last decade, the application of liquid chromatography - high resolution mass spectroscopy (LC-HRMS) has been growing extensively due its ability to identify a wide range of suspect and unknown compounds in environmental samples. However, certain information such as mass accuracy and isotopic pattern of the precursor ion, MS/MS spectra evaluation and retention time plausibility are needed to reach a certain identification confidence. In this context, a comprehensive workflow based on computational tools was developed to understand the retention time behavior of a large number of compounds belonging to emerging contaminants. An extensive dataset was built, containing information for the retention time of 528 and 298 compounds for positive and negative electrospray ionization mode, respectively, to expand the applicability domain of the developed models. Then, the dataset was split into training and test employing k-nearest neighborhood clustering technique so as to build and validate the models' internal and external prediction ability. The best subset of molecular descriptors was selected using genetic algorithms which is based on the evolutionary computations, and could result in representative selection of descriptors. Multiple Linear Regression, Artificial Neural Networks and Support Vector Machines were used to correlate the selected descriptors with the experimental retention times. Several validation techniques were used, including Golbraikh-Tropsha acceptable model criteria's, Euclidean based applicability domain, r2m, concordance correlation coefficient values, to measure the accuracy and precision of the models. The best linear and non-linear models for each dataset were derived and used to predict the retention time of suspect compounds in a wide-scope survey as the evaluation data set. Overall, the proposed workflow was fast, reliable, and less time consuming which can be employed for identification purposes in environmental samples.