Title

Applications of Machine Learning to In Silico Quantification of Chemicals without Analytical Standards

Journal

Journal of Chemical Information and Modeling, 60(6)

ISSN

1549-9596

Authors

Abrahamsson, Dimitri Panagopoulos
Park, June-Soo
Singh, Randolph R
et al.

Publication Date

2020-06-22

DOI

10.1021/acs.jcim.9b01096

Peer reviewed

HHS Public Access

Author manuscript

J Chem Inf Model. Author manuscript; available in PMC 2021 June 22.

Published in final edited form as:

J Chem Inf Model. 2020 June 22; 60(6): 2718–2727. doi:10.1021/acs.jcim.9b01096.

Applications of Machine Learning to *in silico* Quantification of Chemicals without Analytical Standards

Dimitri Panagopoulos Abrahamsson[1,*], **June-Soo Park**[2], **Randolph Singh**[3], **Marina Sirota**[4], **Tracey Woodruff**[1]

[1]Program on Reproductive Health and the Environment, Department of Obstetrics and Gynecology, University of California, San Francisco, CA 94158, USA

[2]Environmental Chemistry Laboratory, California Department of Toxic Substances Control, Berkeley, CA 94710, USA

[3]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg

[4]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA 94158, USA

Abstract

Non-targeted analysis techniques provide a comprehensive approach to analyze environmental and biological samples for nearly all chemicals present in a given sample. One of the main shortcomings of current analytical methods is that they are unable to provide quantitative information about the chemicals in a given sample constituting an important obstacle in understanding environmental fate and human exposure. Herein, we present a machine learning *in silico* quantification method for chemicals analyzed using electrospray ionization (ESI). We considered three different datasets from different instrumental setups: i) capillary electrophoresis electrospray ionization-mass spectrometry (CE-MS) in positive ionization mode (ESI+), ii) liquid chromatography quadrupole time-of-flight mass spectrometry (LC-QTOF/MS) in ESI+ and iii) negative ionization mode (ESI−). We developed and applied two different machine learning algorithms: a random forest (RF) and an artificial neural network (ANN) to predict the relative response factors (RRFs) of different chemicals based on their physicochemical properties. Chemical concentrations can then be calculated by dividing the measured abundance of a chemical, as peak area or peak height, by its corresponding RRF. We evaluated our models and

[*]Corresponding author: Dimitri Panagopoulos Abrahamsson, drdimitripanagopoulos@gmail.com.

Supporting Information

The Supporting Information is available free of charge on the ACS Publication website.

- Information on the design of the algorithms and the optimization of the hyperparameters can be found in ‘Abrahamsson et al. Supporting Information’.
- Chemical names and physicochemical descriptors of the chemicals in the CE-MS ESI+ dataset can be found in ‘Abrahamsson et al. SI spreadsheet 1’.
- Chemical names and physicochemical descriptors of the chemicals in the LC-QTOF/MS datasets can be found in ‘Abrahamsson et al. SI spreadsheet 2’.
- The algorithms are available as Jupyter Notebook files on GitHub (https://github.com/dimitriabrahamsson/expert-octo-spork).

tested their predictive power using 5-fold cross-validation (CV) and y-randomization. Both the RF and the ANN models showed great promise in predicting RRFs. However, the accuracy of the predictions was dependent on the dataset composition and the experimental setup. For the CE-MS ESI+ dataset, the best model predicted measured RRFs with a mean absolute error (MAE) of 0.19 log units and a cross-validation coefficient of determination (Q^2) of 0.84 for the testing set; for the LC-QTOF/MS ESI+ dataset, an MAE of 0.32 and a Q^2 of 0.40; and for the LC-QTOF/MS ESI− dataset, a MAE of 0.50 and a Q^2 of 0.20. Our findings suggest that machine learning algorithms can be used for predicting concentrations of non-targeted chemicals with reasonable uncertainties, especially in ESI+, while the application on ESI− remains a more challenging problem.

Introduction

With recent technological advances in high-resolution mass spectrometry (HRMS) non-targeted analysis (NTA) has arrived at the forefront of analytical chemistry, attracting attention from scientists in analytical, environmental and bioanalytical chemistry as an approach to more comprehensively screen environmental and biological samples. This approach has also attracted the attention of many environmental health scientists and epidemiologists who are studying the human exposome.[1,2] While NTA provides a comprehensive approach to identify potential chemical signatures and exposures, it is limited as it cannot provide quantitative information. This challenge stems from the fact that the number of identified or tentatively identified chemicals is in the range of hundreds to thousands, making it difficult to definitively identify and quantify using chemical standards.[3,4] Additionally, for many identified chemicals there are no available chemical standards on the market.

Concentration estimates are critical when studying environmental fate and human exposure.[5] A common approach to estimate concentrations of non-targeted chemicals is calculating concentrations against an internal standard used for analysis[6] or against multiple internal standards of structurally similar chemicals (quantification markers).[7,8] The main weakness of these approaches is that they do not account for differences in ionization efficiencies across different chemicals. This parameter is often expressed as the relative response factor (RRF) of each compound, which is the ratio of a chemical's abundance (peak area or peak height) to the chemical's concentration in a given sample. Two chemicals at the same concentration can exhibit differences in peak areas spanning several orders of magnitude.[9,10] For example, one chemical may be detected as a small hump barely passing the chromatograph's baseline, while another chemical may be detected as a large peak saturating the detector.[9] Quantification with structurally similar chemicals may give better estimates than a single internal standard, however, even structurally similar compounds can have very different ionization efficiencies[11,12] resulting in large uncertainties in concentration estimates.

In an early study on predicting concentrations in electrospray ionization (ESI) mass spectrometry (MS), Chalcraft et al.[11] developed a predictive model for concentrations of a series of metabolites using the pH-adjusted octanol/water distribution ratio (log *D*), molecular volume (MV), absolute mobility, and effective charge of the analytes. Chalcraft et

al.[11] observed that the two statistically significant parameters in their model were log *D* and MV. In another study, Oss et al.[12] developed a predictive model for concentrations of a series of organic chemicals using the negative logarithm of the acid dissociation constant (p*K*a) in water, p*K*a in acetonitrile, gas phase basicity, molecular mass, molecular area, polar surface and molecular volume of the analytes. Oss et al.[12] observed that p*K*a and MV were the two main driving parameters for describing and predicting the chemicals' ionization efficiency. Obtaining all descriptors and assessing the accuracy of the predictions is a challenging task. In some cases, such as in Chalcraft et al.,[11] obtaining these descriptors requires additional laboratory measurements. In other cases, such as in Oss et al.,[12] obtaining these descriptors requires purchasing licenses to commercial software and/or often the predictions are not accompanied by an estimate of expected uncertainty. One example of open-source software that can be used to calculate physicochemical descriptors is Mordred, which was developed by Moriwaki et al.[13] Mordred can calculate more than 1800 2D and 3D descriptors and it is freely available on GitHub (https://github.com/mordred-descriptor/mordred). Despite Mordred being a very comprehensive tool, the calculated descriptors are not accompanied by estimates of uncertainty and thus it is not possible to know if certain estimates are expected to contain large uncertainties.

Abraham et al.[14,15] proposed a set of descriptors that can be used for predicting physicochemical properties. The predictive equations built with the descriptors proposed by Abraham et al.[14,15] are commonly known as poly-parameter free energy relationships (PP-LFERs). For a given chemical, the descriptors used to build a PP-LFER describe (i) the chemical's ability to engage in London dispersion forces and dipole-induced dipole interactions (E), (ii) the chemical's ability to engage in dipole-induced dipole and dipole-dipole interactions (S), (iii) the chemical's ability to act as a hydrogen-bond donor (A), (iv) the chemical's ability to act as a hydrogen-bond acceptor (B), (v) the chemical's McGowan molecular volume (V), and (vi) the chemical's hexadecane/air partition ratio (L).[14,15] The descriptors can be downloaded from the database of the Helmholtz Centre for Environmental Research (Helmholtz Centrum für Umweltforschung; UFZ; https://www.ufz.de/lserd/) called UFZ-LSER.[16] The UFZ-LSER database contains both experimentally determined and predicted descriptors. The predicted descriptors are calculated based on the structural features of a given chemical. Initially, these descriptors were meant to be used in predictions of partition ratios,[17,18] such as the partition ratio between organic carbon and water (K_{oc}) or the partition ratio between octanol and water (K_{ow}), but they have also been successfully used in predictions of other physicochemical properties, such as the Setschenow (or salting-out) constant (K^{s}).[19,20] It is important to note that the predicted Abraham descriptors are always accompanied by estimated uncertainties and warnings with regards to the expected quality of the descriptors. These estimates and warnings are given by the UFZ-LSER database together with the Abraham descriptors summarized in spreadsheet 1 in SI, sheet: "readme for Abraham descriptors."

One of the shortcomings of PP-LFERs is their limited ability to cover chemicals from classes that were not included in the training set that the PP-LFERs were built on. For example, previous studies[17,21] have shown some evidence that PP-LFERs that did not include organosilicon compounds in their training sets failed to accurately describe the physicochemical properties of these compounds.

Machine learning technologies, such as random forest (RF), support-vector machine (SVM) and artificial neural networks (ANN), have been successful in image analysis and voice recognition[22] and have become particularly popular with the development of Google Brain[23] and TensorFlow[24] by Google. These technologies have also found applications in various areas of chemistry, especially in the areas of physical chemistry and cheminformatics. Some examples of these applications are predictions of physicochemical properties using quantum chemical activity/property relationships (QSAR/QSPR),[25-27] predictions of quantum mechanical properties of molecules [28,29] and in cheminformatics for drug discovery.[30] One important difference between machine learning and multi-linear regressions (MLRs), such as PP-LFERs, is that while MLRs assume that chemical structure and activity are directly related through a set number of descriptors, machine learning allows for that relationship to be formed through a series of nodes and additional descriptors created during the model training.

Our study aims to develop machine learning algorithms using the RF and ANN methodologies for predictions of RRFs, which can then be used to calculate concentrations of chemicals without analytical standards.

Materials and Methods

Data

We considered three different datasets for the purposes of this study: (1) a dataset with 57 polar endogenous compounds analyzed with Capillary Electrophoresis - Electrospray Ionization - Mass Spectrometry (CE-MS) in positive ionization mode (ESI+) from the study of Chalcraft et al.,[11] (2) 517 chemicals analyzed with Liquid Chromatography – Quadrupole – Time-Of-Flight / Mass Spectrometry (LC-QTOF/MS) in ESI+ and (3) 254 chemicals analyzed with LC-QTOF/MS in negative ionization mode (ESI−) from the study of Sobus et al.[31] The methods for the sample preparation and analysis are described in detail in the studies of Chalcraft et al.[11] and Sobus et al.[31]

In Chalcraft et al.,[11] 57 polar endogenous compounds (amino acids, amines, peptides, acylcarnitines, and nucleosides) (spreadsheet 1 in SI) and 1 internal standard (total n = 58) were diluted in methanol at six different concentration levels (2, 8, 16, 30, 50, 100 μM) with 1.4M formic acid as a background electrolyte and were analyzed with Capillary Electrophoresis - Electrospray Ionization - Mass Spectrometry (CE-MS ESI+). All analyses were conducted in ESI+. The RRF for each chemical was represented by the slope of each calibration curve after fitting a linear regression through the calibration points. It is important to note that all calibration curves in the study of Chalcraft et al.[7] were made by using the peak areas of the chemicals normalized to the areas of the internal standard. The ratio of the peak area of the chemical to the peak area of the internal standard is presented in that paper as the relative ion response (RIR). The RIR is related to concentration via RRF:

$$RRF = \frac{RIR}{C} \quad (1)$$

where, C is the concentration of the chemical.

In Sobus et al.,[31] the chemicals were initially diluted in dimethyl sulfoxide at approximately 0.05 mM for each chemical. Dilutions were then prepared in methanol and a buffer solution of HPLC water containing 2 mM ammonium formate at a ratio of 1:3 methanol:buffer yielding nominal concentrations of approximately 0.5, 0.1 and 0.02 μM. Aliquots of 400 μL of the diluted samples were then spiked with 10 μL of a solution containing stable isotope-labeled tracers in methanol at 1 ng/μL ($^{13}C_6$-methyl paraben, $^{13}C_6$-butyl paraben, $^{13}C_4$-periluorooctanoic acid, $^{13}C_4$ $^{15}N_2$-Fipronil, $^{13}C_4$ $^{15}N_2$-Fipronil sulfone, $^{13}C_5$-perfluorononanoic acid, $^{13}C_4$-perfluooctanesulfonic acid, $^{13}C_2$-periluorodecanoic acid, $^{13}C_3$-atrazine, D_3-thiamethoxam, and D_4-pyriproxyfen). Each sample was then analyzed in triplicate injections using LC-QTOF/MS in both ESI+ and ESI−. The RRFs for the chemicals from this study were calculated by dividing the chemicals' peak areas by the chemicals' concentration:

$$RRF = \frac{A}{C} \quad (2)$$

where, A is the abundance (peak areas) of the chemical.

Data Analysis

Our workflow is presented schematically as a diagram in Figure 1. A detailed description of the data collection of each dataset is presented in the sections below.

CE-MS ESI+ dataset

We collected Abraham descriptors for all chemicals from the database of the Helmholtz Centre for Environmental Research (UFZ-LSER, https://www.ufz.de/lserd/)[16] and first constructed a PP-LFER, which we compared to an MLR constructed with the physicochemical descriptors measured in Chalcraft et al.[11] We then constructed an RF using the scikit-learn[32] platform and an ANN using the Tensorflow[24] platform. The scripting was done in Python.[33] For the RF and the ANN, we examined three modeling scenarios; in scenario 1 (called "RF" and "ANN") we used the same physicochemical descriptors as in the PP-LFERs; in scenario 2 (called "RF [+]" and "ANN [+]") we expanded on the physicochemical properties by adding the ChemmineR[34,35] physicochemical descriptors from the online tool ChemmineR developed by the University of California, Riverside;[34,35] and in scenario 3, we repeated the calculations of scenario 1 after removing the Abraham descriptors that carried warnings of high expected uncertainty. Scenario 1 helped us to directly compare the predictive power of PP-LFERs to that of RFs and ANNs by using the exact same physicochemical descriptors. Scenario 2 provided an insight as to whether we could improve the performance of the RF and ANN models by adding more physicochemical descriptors. The descriptors derived by ChemmineR are structural descriptors that outline how many atoms of a specific element (e.g., number of carbons, C) and how many specific functional groups (e.g. number of RCOOH groups) are present in a molecule. All the ChemmineR descriptors for all chemicals can be found in spreadsheet 1 in SI. Finally, scenario 3 helped us assess whether the expected errors in the Abraham descriptors propagate into errors in the predictions of RRFs.

LC-QTOF/MS ESI+ and ESI− datasets

We collected Abraham and ChemmineR descriptors for all chemicals in the dataset. In addition, we also collected Mordred descriptors from the Mordred Python package developed by Moriwaki et al.[13] We constructed an RF and an ANN as described for the Chalcraft et al. dataset and we tested four different scenarios. In scenario 1, called “RF” and “ANN,” we built a model using only the Abraham descriptors. In scenario 2, called “RF[+]” and “ANN[+],” we built a model using the Abraham descriptors with the ChemmineR descriptors. In scenario 3, called “RF[-]” and “ANN[-],” we repeated the calculations in scenario 2 after removing the Abraham descriptors that are expected to contain large uncertainties based on the warnings given by the UFZ-LSER database. In scenario 4, (called “RF[m]” and “ANN[m]”), we constructed a model using the Mordred descriptors instead of the Abraham and ChemmineR descriptors.

Model Validation

To test the models for their predictive power and to control for over-fitting we applied a 5-fold cross-validation (CV). We randomly divided the dataset into training and testing sets following an 80/20 split. The process was repeated five times to ensure that the majority of the chemicals were included once in the training set and once in the testing set. This approach is expected to provide a more representative picture of the accuracy of the predictions compared to an 1-fold CV. In each one of the five replicates, the constructed models were used to predict the RRFs of the chemicals from the group that was left out of the training set (testing set). We compiled the results from all five attempts into one dataset and calculated the CV mean absolute error (MAE_{CV}) and the CV coefficient of determination (Q^2). MAE_{CV} was calculated as:

$$MAE_{CV} = \frac{\sum_{i=1}^{n} | y_{exp} - y_{pred} |}{n} \quad (3)$$

where, y_{exp} is the experimental parameter, y_{pred} is the predicted parameter and n is the total number of observations.

In addition to the 5-fold CV, we conducted a y-randomization analysis for the best performing models of each dataset to ensure that the selected descriptors have some predictive potential and that the models are not predicting randomly.[36] For the y-randomization analysis, we kept the descriptors (X variable) as they were, and we randomized the experimental RRFs (y variable). We then divided the dataset into training and testing sets following an 80/20 split as in the 5-fold CV, and then repeated the process 5 times. We calculated the MAE and Q^2 for both the training and the testing sets and compared the results to those of the original CV analysis.

The model design and compilation are described in detail in Text S1. The constructed RFs and ANNs with all the parameters can be found in the Jupyter Notebook files uploaded on GitHub (https://github.com/dimitriabrahamsson/expertocto-spork).

Results and Discussion

CE-MS ESI+ dataset

The RF and the ANN (Fig. 2) showed the strongest predictive power of all models both in terms of Q^2 and MAE_{CV} (Fig. 2). When comparing the Q^2 values of the testing sets of the ANN and RF to the Q^2 values of their training sets (Fig. 2), we observed a good agreement between the two datasets indicating that there was no substantial over-fitting for either the ANN or the RF model. The addition of the ChemmineR descriptors in RF and ANN (RF[+] and ANN[+]) slightly increased the errors in the predictions for the testing sets of both the RF[+] and the ANN[+] (Fig. 3) and for the training set of ANN[+] (Fig. 3). This observation suggests that some of the ChemmineR descriptors may not be useful in describing the process of electrospray ionization of the chemicals in this dataset and introduce small errors in the predictions (Fig. 3). However, this observation should be confirmed with larger and more structurally diverse datasets.

The PP-LFERs displayed predictive power similar to that of the MLRs (Fig. 2; Tables S1 and S2). This is particularly important because if we can use the predicted Abraham descriptors instead of the experimentally determined descriptors of Chalcraft et al.,[11] then this would simplify and accelerate the data collection for concentration predictions of chemicals in large datasets from NTA. As mentioned earlier, the predicted Abraham descriptors are always accompanied by estimated uncertainties and warnings with regards to the expected quality of the descriptors. In the CE-MS ESI+ dataset, there were seven chemicals, for which the predicted Abraham descriptors from the UFZ-LSER database were expected to contain large uncertainties (spreadsheet 1 in SI, sheets: “Abraham descriptors” and “readme for Abraham descriptors”). The chemicals were L-lysine, cystathionine, oxidized glutathione, oxytetracycline, L-ornithine, L-citrulline and L-arginine (spreadsheet 1 in SI, sheet: “Abraham descriptors”; chemicals shown in red font). According to the explanation given by the researchers who compiled the UFZ-LSER database,[16] the predicted Abraham descriptors are expected to be inaccurate because the chemicals are outside the domain of applicability of the UFZ-LSER models. This is often the case when the chemicals are structurally very distant from the chemicals in the training set of the UFZ-LSER models. A detailed explanation of how these warnings are generated is given in spreadsheet 1 in SI (sheet: “readme for Abraham descriptors”). After removing these chemicals and repeating the CV, the predictive power of the PP-LFER improved dramatically (Fig. 4; Tables S2 and S3). The Q^2 of the testing set increased from 0.23 to 0.75 and the MAE decreased from 0.51 to 0.32. These findings indicate that PP-LFERs are capable of producing accurate predictions of RRFs as long as the physicochemical descriptors, which the PP-LFERs are built on, do not contain any substantial uncertainties.

In order to understand if the addition of these chemicals to the dataset had a negative impact on the predictions of the RF and ANN models, we repeated the 5-fold CV with the reduced dataset (Fig. 4). Removing these chemicals did not seem to improve the predictions of the RF and ANN models. Interestingly enough, it slightly worsened the predictions of both the RF and the ANN models. However, the change seems to be minimal (Fig. 4). Perhaps this

observation is not connected to the removal of chemicals per se, but to a substantial reduction in the dataset (12%) and a reduction in structural variability.

Comparing our results to the results of Chalcraft et al.[11] we observed that the MLR model from our study showed similar predictive power to the MLR model built in Chalcraft et al.,[11] who reported a mean absolute error of 40% for the predictions of the testing set. This value is very close to the MAE_{CV} we observed in our predictions of RRFs for the testing set using our MLR model, 0.42 log units (Fig. 2). Furthermore, our RF and ANN models showed a substantially improved accuracy compared to the MLR models built here and in Chalcraft et al.[11] with the MAE_{CV} of RF and ANN (Fig. 2) being approximately half that of the MLR models (MAE_{CV} RF: 0.19, MAE_{CV} ANN: 0.19 vs MAE_{CV} MLR: 0.40).

LC-QTOF/MS ESI+ dataset

The predictions for the LC-QTOF/MS ESI+ dataset (Fig. 5) showed a decrease in accuracy compared to the predictions made for the CE-MS ESI+ dataset (Fig. 3). This observation is somewhat expected considering that the LC-QTOF/MS ESI+ dataset is substantially larger in size and a lot more structurally diverse than the CE-MS ESI+ dataset. While the CE-MS ESI+ dataset contains only 57 endogenous compounds many of which are structurally similar, the LC-QTOF/MS dataset contains 517 chemicals of diverse sources and structures. The RF models performed comparably well to the ANN models, with the RFs showing slightly better performance than the ANNs (Fig. 5). Contrary to the findings for the CE-MS ESI+ dataset, the addition of the ChemmineR descriptors to the LC-QTOF/MS datasets resulted in a slight improvement of the predictions increasing the Q^2 and decreasing the MAE of the testing sets (Fig. 5A-B and E-F). This finding suggests that the addition of these descriptors might be of value in larger and structurally more diverse datasets compared to the CE-MS ESI+ dataset. In the LC-QTOF/MS ESI+ dataset there were 69 chemicals for which at least one descriptor was expected to contain high uncertainties according to the warnings given by the UFZ-LSER database (SI spreadsheet 2, sheet “Abraham descriptors ESI+”). Contrary to the findings from the CE-MS ESI+ dataset, removing these chemicals slightly improved the predictions of the models for the testing set increasing the Q^2 and decreasing the MAE for both the RF and the ANN (Fig. 5A-C and E-G). However, also in this case, the effect of that change seems to be minimal. The RF built with the Mordred descriptors, RF [m], showed slightly better performance compared to RF, RF [+] and RF [-]. However, in the case of the ANN models, the ANN built with the Mordred descriptors, ANN [m], produced slightly less accurate predictions compared to the ANN [+] and ANN [-] scenarios (Fig. 5). Comparing the RF [m] with the ANN [m] models (Fig. 5), the RF [m] performed slightly better in terms of Q^2. However, in terms of prediction errors, the MAE_{CV} for the testing set of the two models were very similar (Fig. 5D and H).

Comparing our findings from the LC-QTOF/MS ESI+ dataset to the study of Chalcraft et al.[11], we observed that the MAE_{CV} of the best machine learning model, RF [m] (Fig. 5D), was 0.08 log units lower than that of the MLR of Chalcraft et al.[11] It is also important to note that the LC-QTOF/MS ESI+ contained 518 chemicals whereas the Chalcraft et al.[11] dataset contained only 57 chemicals. This comparison, together with the findings for the CE-MS

dataset, demonstrates the greater potential of RFs and ANNs compared to MLRs in predictions of RRFs.

LC-QTOF/MS ESI− dataset

The predictions for the LC-QTOF/MS ESI− dataset (Fig. 6) were overall less accurate than the predictions for LC-QTOF/MS ESI+ dataset (Fig. 5). Although the exact reason behind this observation remains unknown, this finding suggests that the process of negative ionization could not be effectively described with the physicochemical descriptors that were examined in this study. Overall, the RF models made slightly more accurate predictions for the testing set compared to the ANN models increasing the Q^2 and decreasing the MAE_{CV} (Fig. 6). As in the case of the LC-TOF/MS ESI+ dataset, the addition of the ChemmineR descriptors showed a slight improvement in the predictions increasing the Q^2 and decreasing the MAE_{CV} for both the RF and ANN models (Fig. 6A-B and E-F). In the LC-QTOF/MS ESI− dataset, there were 39 chemicals, for which at least one descriptor was expected to contain high uncertainties (SI spreadsheet 2, sheet “Abraham descriptors ESI-”). Removing these chemicals from the dataset resulted in slightly improved predictions for the RF model, RF [-], but slightly worse predictions for the ANN [-] (Fig. 6A-C and E-G). The models built with the Mordred descriptors, RF [m] and ANN [m], showed the best performance compared to the other models with the RF [m] performing better than ANN [m]. However, in terms of prediction errors, the MAE_{CV} of the two models were very similar (Fig. 6). These findings are in good agreement with the observations we made for the LC-QTOF/MS ESI+ dataset.

We compared our findings from the LC-QTOF/MS ESI− dataset to the study of Kruve et al. [37], which developed an MLR for predicting ionization efficiencies of chemicals analyzed with an ion trap mass spectrometer operated in ESI-. We observed that the MAE_{CV} of our best model, RF [m], was 0.02 log units higher than error reported for the testing set of the MLR model of Kruve et al.[37] (RMSE = 0.48 log units). It is important to note here that the LC-QTOF/MS ESI− contained 254 chemicals, whereas the dataset of Kruve et al.[37] contained only 63 chemicals, which were structurally similar comprising three chemical classes: benzoic acids, phenols, and salicylic acids. This observation confirms our previous conclusions about the greater potential of RFs and ANNs compared to MLRs in predictions of RRFs.

Y-randomization

For the y-randomization analysis, we selected the best performing models from the three datasets (Fig. 7). In all three cases, y-randomization resulted in substantially lower Q^2 values and increased the MAE in the testing set for the majority of the RF and the ANN models (Fig. 5-7). This observation supported our findings from the 5-fold CV analysis and confirmed that the selected physicochemical descriptors have predictive importance and that the RF and ANN models do not predict randomly. The largest differences were observed for the CE-MS ESI+ dataset, where the MAE of the testing set increased from 0.19 to 0.56 for the RF model and from 0.19 to 0.61 for the ANN model (Fig. 7 and 4). Interestingly enough, even for the negative ionization data, LC-QTOF/MS ESI-, where the predictions carried

substantial errors, randomizing the y-variable increased the MAE of the testing set from 0.54 to 0.63 (Fig 7H and 6H).

Limitations and future considerations

To our knowledge, this is the first study to evaluate applications of machine learning to in silico quantification of chemicals when using ESI MS. One limitation of our approach that needs to be taken into consideration is that non-targeted analysis is conducted using various types of instruments such as CE-MS, QTOF/MS, ion trap MS, and Orbitrap MS, and the ionization efficiencies of chemicals may vary depending on the instrument used for analysis.[38] Additionally, different analytical methods and different sample matrices, such as blood, urine or water, are also known to influence ionization efficiency.[39-41]

For these reasons, when applying our methodology, it is important to note that safe predictions can only be made for the specific instrument type and analytical method used when building the models. Thus, validation of the constructed models with an external dataset from a different instrument and/or a different method is not expected to provide useful conclusions. In future studies, we aim to test our quantitative models using various datasets of structurally diverse chemicals from various instruments and various analytical methods to expand on the applicability of our models.

One shortcoming when working with machine learning algorithms is that it is often difficult to interpret features from the developed algorithm making it hard to draw useful conclusions about the influence of each feature on the predictions. Nonetheless, machine learning algorithms have shown great promise in providing concentration estimates for chemicals when it is not feasible to quantify these using analytical standards or when analytical standards are not commercially available. We hope with our study that we can make a contribution to the ongoing discussion about non-targeted analysis and the in silico quantification of chemicals and to help bridge the gap between non-targeted analysis, environmental fate, and human exposure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This project was funded by the NIH/NIEHS (P01ES022841, R01ES027051, K01LM012381) and by the US EPA funding: US EPA (RD 83543301 and 83564301). We would like to thank Jon Sobus for assisting with the data collection for the LC-QTOF/MS datasets and Courtney Cooper for writing assistance, language editing and proofreading.

References

(1). The Exposome; Elsevier, 2014 10.1016/C2013-0-06870-3.

(2). Wild CP Complementing the Genome with an “Exposome”: The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. Cancer Epidemiol. Prev. Biomark 2005, 14 (8), 1847–1850. 10.1158/1055-9965.EPI-05-0456.

(3). Rager JE; Strynar MJ; Liang S; McMahen RL; Richard AM; Grulke CM; Wambaugh JF; Isaacs KK; Judson R; Williams AJ; Sobus JR Linking High Resolution Mass Spectrometry Data with

Exposure and Toxicity Forecasts to Advance High-Throughput Environmental Monitoring. Environ. Int 2016, 88, 269–280. 10.1016/j.envint.2015.12.008. [PubMed: 26812473]

(4). Moschet C; Anumol T; Lew BM; Bennett DH; Young TM Household Dust as a Repository of Chemical Accumulation: New Insights from a Comprehensive High-Resolution Mass Spectrometric Study. Environ. Sci. Technol 2018, 52 (5), 2878–2887. 10.1021/acs.est.7b05767. [PubMed: 29437387]

(5). Wambaugh JF; Setzer RW; Reif DM; Gangwal S; Mitchell-Blackwood J; Arnot JA; Joliet O; Frame A; Rabinowitz J; Knudsen TB; Judson RS; Egeghy P; Vallero D; Cohen Hubal EA High-Throughput Models for Exposure-Based Chemical Prioritization in the ExpoCast Project. Environ. Sci. Technol 2013, 47 (15), 8479–8488. 10.1021/es400482g. [PubMed: 23758710]

(6). Phillips KA; Yau A; Favela KA; Isaacs KK; McEachran A; Grulke C; Richard AM; Williams AJ; Sobus JR; Thomas RS; Wambaugh JF Suspect Screening Analysis of Chemicals in Consumer Products. Environ. Sci. Technol 2018, 52 (5), 3125–3135. 10.1021/acs.est.7b04781. [PubMed: 29405058]

(7). Pieke EN; Granby K; Trier X; Smedsgaard J A Framework to Estimate Concentrations of Potentially Unknown Substances by Semi-Quantification in Liquid Chromatography Electrospray Ionization Mass Spectrometry. Anal. Chim. Acta 2017, 975, 30–41. 10.1016/j.aca.2017.03.054. [PubMed: 28552304]

(8). Go Y-M; Walker DI; Liang Y; Uppal K; Soltow QA; Tran V; Strobel F; Quyyumi AA; Ziegler TR; Pennell KD; Miller GW; Jones DP Reference Standardization for Mass Spectrometry and High-Resolution Metabolomics Applications to Exposome Research. Toxicol. Sci. Off. J. Soc. Toxicol 2015, 148 (2), 531–543. 10.1093/toxsci/kfv198.

(9). Kruve A Strategies for Drawing Quantitative Conclusions from Non-Targeted Liquid Chromatography High-Resolution Mass Spectrometry Analysis. Anal. Chem 2020 10.1021/acs.analchem.9b03481.

(10). Kebarle P; Tang L From Ions in Solution to Ions in the Gas Phase - the Mechanism of Electrospray Mass Spectrometry. Anal. Chem 1993, 65 (22), 972A–986A. 10.1021/ac00070a001.

(11). Chalcraft KR; Lee R; Mills C; Britz-McKibbin P Virtual Quantification of Metabolites by Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry: Predicting Ionization Efficiency without Chemical Standards. Anal. Chem 2009, 81 (7), 2506–2515. 10.1021/ac802272u. [PubMed: 19275147]

(12). Oss M; Kruve A; Herodes K; Leito I Electrospray Ionization Efficiency Scale of Organic Compounds. Anal. Chem 2010, 82 (7), 2865–2872. 10.1021/ac902856t. [PubMed: 20218595]

(13). Moriwaki H; Tian Y-S; Kawashita N; Takagi T Mordred: A Molecular Descriptor Calculator. J. Cheminformatics 2018, 10 (1), 4 10.1186/s13321-018-0258-y.

(14). Abraham MH Scales of Solute Hydrogen-Bonding: Their Construction and Application to Physicochemical and Biochemical Processes. Chem. Soc. Rev 1993, 22 (2), 73–83. 10.1039/CS9932200073.

(15). Abraham MH; Ibrahim A; Zissimos AM Determination of Sets of Solute Descriptors from Chromatographic Measurements. J. Chromatogr. A 2004, 1037 (1), 29–47. 10.1016/j.chroma.2003.12.004. [PubMed: 15214659]

(16). UFZ - LSER Database https://www.ufz.de/index.php?en=31698&contentonly=1&m=0&lserd_data[mvc]=Public/start (accessed Feb 17, 2020).

(17). Panagopoulos D; Jahnke A; Kierkegaard A; MacLeod M Organic Carbon/Water and Dissolved Organic Carbon/Water Partitioning of Cyclic Volatile Methylsiloxanes: Measurements and Polyparameter Linear Free Energy Relationships. Environ. Sci. Technol 2015, 49 (20), 12161–12168. 10.1021/acs.est.5b02483. [PubMed: 26371969]

(18). Goss K-U Predicting the Equilibrium Partitioning of Organic Compounds Using Just One Linear Solvation Energy Relationship (LSER). Fluid Phase Equilibria 2005, 233 (1), 19–22. 10.1016/j.fluid.2005.04.006.

(19). Endo S; Pfennigsdorff A; Goss K-U Salting-Out Effect in Aqueous NaCl Solutions: Trends with Size and Polarity of Solute Molecules. Environ. Sci. Technol 2012, 46 (3), 1496–1503. 10.1021/es203183z. [PubMed: 22191628]

(20). Panagopoulos D; Kierkegaard A; Jahnke A; MacLeod M Evaluating the Salting-Out Effect on the Organic Carbon/Water Partition Ratios (KOC and KDOC) of Linear and Cyclic Volatile Methylsiloxanes: Measurements and Polyparameter Linear Free Energy Relationships. J. Chem. Eng. Data 2016, 61 (9), 3098–3108. 10.1021/acs.jced.6b00196.

(21). Endo S; Goss K-U Predicting Partition Coefficients of Polyfluorinated and Organosilicon Compounds Using Polyparameter Linear Free Energy Relationships (PP-LFERs). Environ. Sci. Technol 2014, 48 (5), 2776–2784. 10.1021/es405091h. [PubMed: 24491038]

(22). Software, D. 9 Applications of Machine Learning from Day-to-Day Life https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0 (accessed Feb 17, 2020).

(23). Google AI https://ai.google/ (accessed Feb 17, 2020).

(24). TensorFlow https://www.tensorflow.org/ (accessed Feb 17, 2020).

(25). Soto AJ; Cecchini RL; Vazquez GE; Ponzoni I Multi-Objective Feature Selection in QSAR Using a Machine Learning Approach. QSAR Comb. Sci 2009, 28 (11–12), 1509–1523. 10.1002/qsar.200960053.

(26). Chinta S; Rengaswamy R Machine Learning Derived Quantitative Structure Property Relationship (QSPR) to Predict Drug Solubility in Binary Solvent Systems. Ind. Eng. Chem. Res 2019, 58 (8), 3082–3092. 10.1021/acs.iecr.8b04584.

(27). Wang Z; Su Y; Shen W; Jin S; Clark JH; Ren J; Zhang X Predictive Deep Learning Models for Environmental Properties: The Direct Calculation of Octanol–Water Partition Coefficients from Molecular Graphs. Green Chem. 2019, 21 (16), 4555–4565. 10.1039/C9GC01968E.

(28). Rupp M; Ramakrishnan R; von Lilienfeld OA Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. J. Phys. Chem. Lett 2015, 6 (16), 3309–3313. 10.1021/acs.jpclett.5b01456.

(29). Krems RV Bayesian Machine Learning for Quantum Molecular Dynamics. Phys. Chem. Chem. Phys 2019, 21 (25), 13392–13410. 10.1039/C9CP01883B. [PubMed: 31165115]

(30). Lo Y-C; Rensi SE; Torng W; Altman RB Machine Learning in Chemoinformatics and Drug Discovery. Drug Discov. Today 2018, 23 (8), 1538–1546. 10.1016/j.drudis.2018.05.010. [PubMed: 29750902]

(31). Sobus JR; Grossman JN; Chao A; Singh R; Williams AJ; Grulke CM; Richard AM; Newton SR; McEachran AD; Ulrich EM Using Prepared Mixtures of ToxCast Chemicals to Evaluate Non-Targeted Analysis (NTA) Method Performance. Anal. Bioanal. Chem 2019, 411 (4), 835–851. 10.1007/s00216-018-1526-4. [PubMed: 30612177]

(32). scikit-learn: machine learning in Python — scikit-learn 0.22.1 documentation https://scikit-learn.org/stable/ (accessed Feb 20, 2020).

(33). Welcome to Python.org https://www.python.org/ (accessed Feb 20, 2020).

(34). ChemMine Tools https://chemminetools.ucr.edu/ (accessed Feb 17, 2020).

(35). Cao Y; Charisi A; Cheng L-C; Jiang T; Girke T ChemmineR: A Compound Mining Framework for R. Bioinformatics 2008, 24 (15), 1733–1734. 10.1093/bioinformatics/btn307. [PubMed: 18596077]

(36). Rücker C; Rücker G; Meringer M Y-Randomization and Its Variants in QSPR/QSAR. J. Chem. Inf. Model 2007, 47 (6), 2345–2357. 10.1021/ci700157b. [PubMed: 17880194]

(37). Kruve A; Kaupmees K; Liigand J; Leito I Negative Electrospray Ionization via Deprotonation: Predicting the Ionization Efficiency. Anal. Chem 2014, 86 (10), 4822–4830. 10.1021/ac404066v. [PubMed: 24731109]

(38). Liigand J; Kruve A; Liigand P; Laaniste A; Girod M; Antoine R; Leito I Transferability of the Electrospray Ionization Efficiency Scale between Different Instruments. J. Am. Soc. Mass Spectrom 2015, 26 (11), 1923–1930. 10.1007/s13361-015-1219-6. [PubMed: 26246121]

(39). Benijts T; Dams R; Lambert W; De Leenheer A Countering Matrix Effects in Environmental Liquid Chromatography–Electrospray Ionization Tandem Mass Spectrometry Water Analysis for Endocrine Disrupting Chemicals. J. Chromatogr. A 2004, 1029 (1), 153–159. 10.1016/j.chroma.2003.12.022. [PubMed: 15032360]

(40). Schuhmacher J; Zimmer D; Tesche F; Pickard V Matrix Effects during Analysis of Plasma Samples by Electrospray and Atmospheric Pressure Chemical Ionization Mass Spectrometry:

Practical Approaches to Their Elimination. Rapid Commun. Mass Spectrom 2003, 17 (17), 1950–1957. 10.1002/rcm.1139. [PubMed: 12913858]

(41). Taylor PJ Matrix Effects: The Achilles Heel of Quantitative High-Performance Liquid Chromatography–Electrospray–Tandem Mass Spectrometry. Clin. Biochem 2005, 38 (4), 328–334. 10.1016/j.clinbiochem.2004.11.007. [PubMed: 15766734]

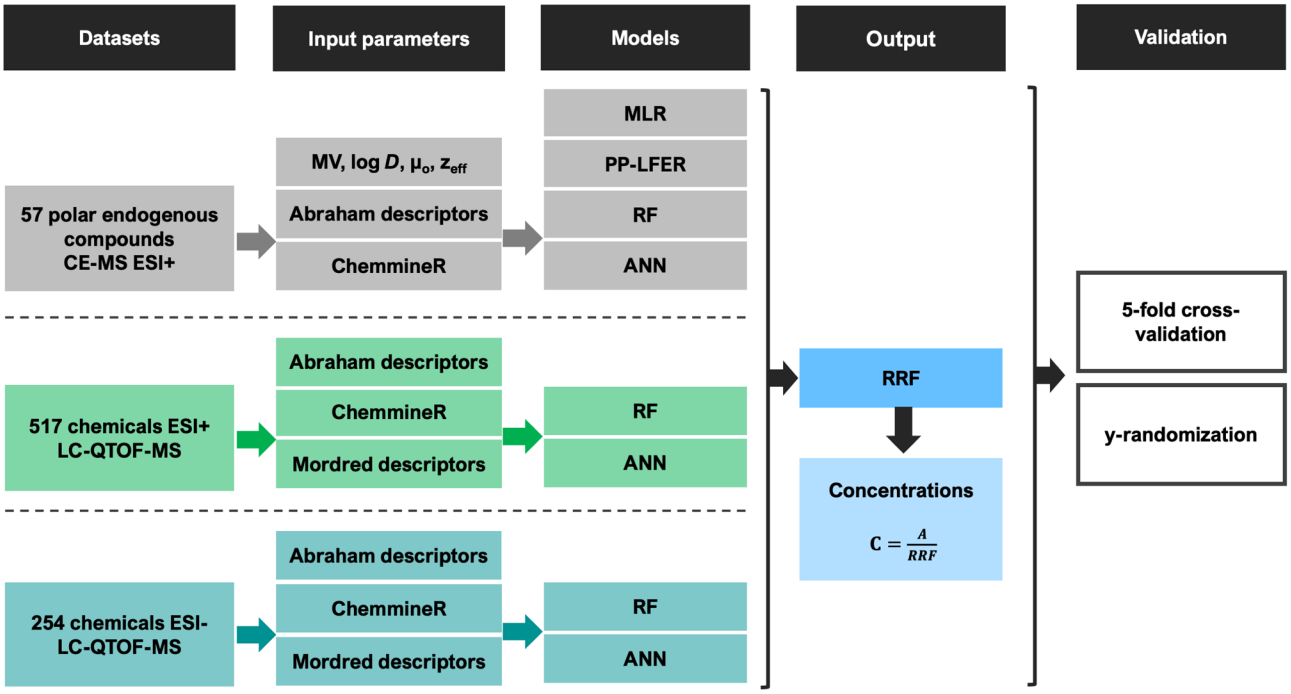

Figure 1:

Workflow diagram for the collection of the experimental data from three independent datasets (CE-MS ESI+, LC-QTOF/MS ESI+, and LC-QTOF/MS ESI−), the physicochemical descriptors that were used as input parameters for the models (Chalcraft descriptors: MV, log D, μ_o, zeff, Abraham descriptors, ChemmineR descriptors, and Mordred descriptors) the models used for each dataset (MLR, PP-LFER, RF, and ANN), the output of the model (RRF) and the model validation (5-fold cross-validation and y-randomization).

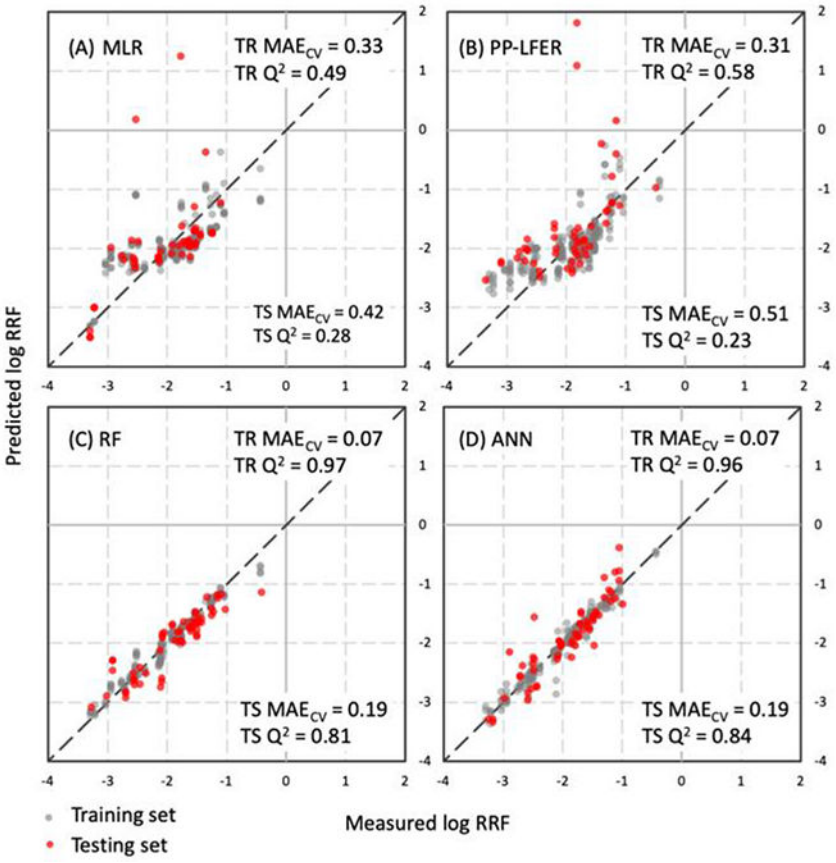

Figure 2:

Results of the 5-fold CV for the CE-MS ESI+ dataset. (A): MLRs built with the experimentally determined physicochemical descriptors from Chalcraft et al.,[11] (B): PP-LFER, (C): RF and (D): ANN built using the Abraham physicochemical descriptors.[9,10] The 5-fold CV was conducted by randomly dividing the dataset into training and testing sets following an 80/20 split, building the model using the training set and testing it using the testing set. The process was repeated five times for each model and the models were assessed based on the CV mean absolute error (MAE_{CV}) of the predictions and the CV coefficient of determination (Q^2).

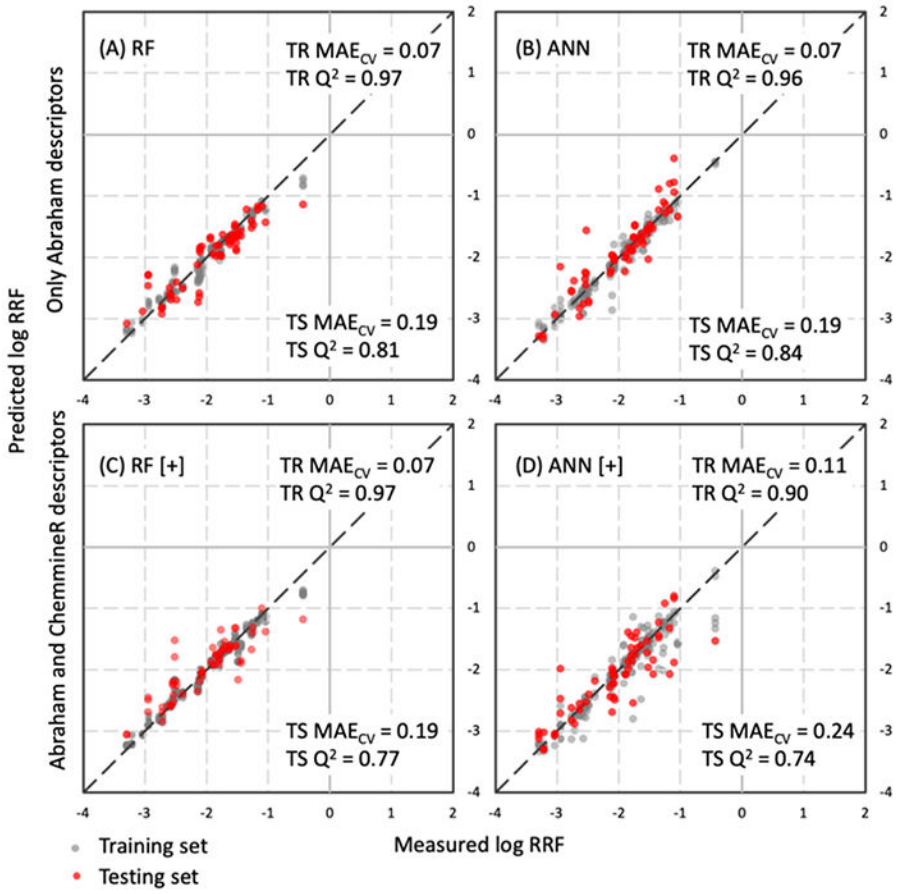

Figure 3:

Results of the 5-fold CV for the CE-MS ESI+ dataset. A and B: Training and testing sets of RF and ANN models built using only the Abraham physicochemical descriptors. C and D: RF [+] and ANN [+] models built using both the Abraham and the ChemmineR descriptors. The 5-fold CV was conducted by randomly dividing the dataset into training and testing sets following an 80/20 split, building the model using the training set and testing it using the testing set. The process was repeated five times for each model and the models were assessed based on the CV mean absolute error (MAE_{CV}) of the predictions and the CV coefficient of determination (Q^2).

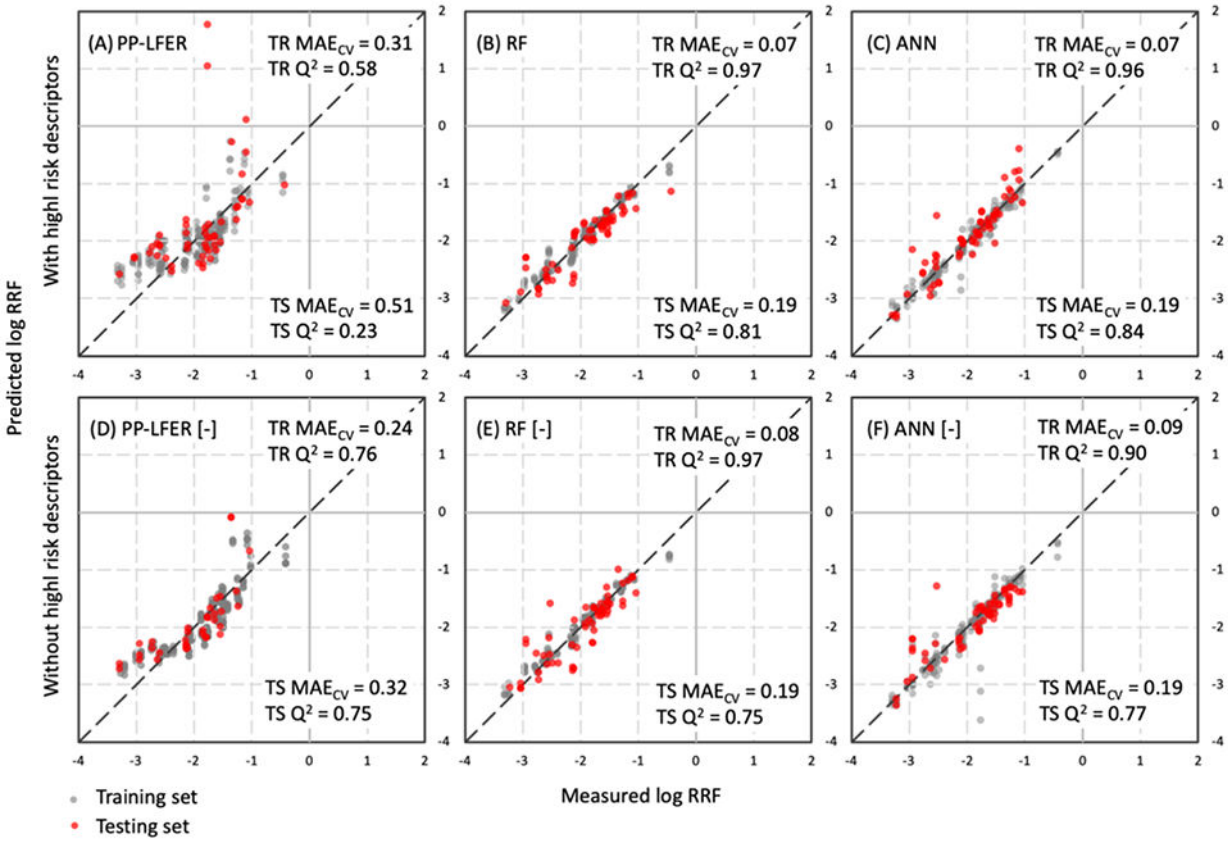

Figure 4:

Results of the 5-fold CV for the CE-MS ESI+ dataset. A, B and C: Training and testing sets of the models built using the Abraham physicochemical descriptors before removing the compounds (n=7) for which at least one descriptor was expected to be inaccurate. D, E, and F: Training and testing sets after removing the seven compounds. The 5-fold CV was conducted by randomly dividing the dataset into training and testing set following an 80/20 split, building the model using the training set and testing it using the testing set. The process was repeated five times for each model and the models were assessed based on the CV mean absolute error (MAE_{CV}) of the predictions and the CV coefficient of determination (Q^2).

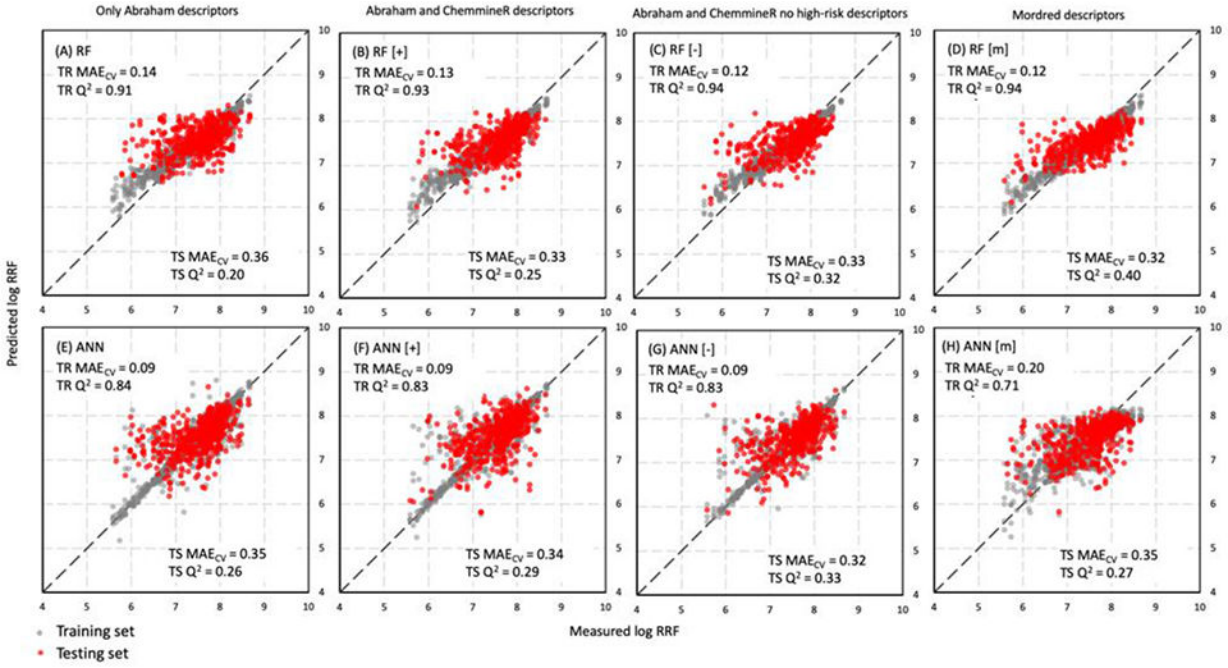

Figure 5:

5-fold CV analysis for the LC-QTOF/MS ESI+ dataset for scenario 1 (only Abraham descriptors; A and E), scenario 2 (Abraham and ChemmineR descriptors; B and F), scenario 3 (Abraham and ChemmineR without high-risk descriptors; C and G) and scenario 4 (Mordred descriptors; D and H). The top panels (A-D) present the calculations for the RF models and the bottom panels (E-H) present the calculations for the ANN models. The 5-fold CV was conducted by randomly dividing the dataset into training and testing set following an 80/20 split, building the model using the training set and testing it using the testing set. The process was repeated five times for each model and the models were assessed based on the CV mean absolute error (MAE_{CV}) of the predictions and the CV coefficient of determination (Q^2).

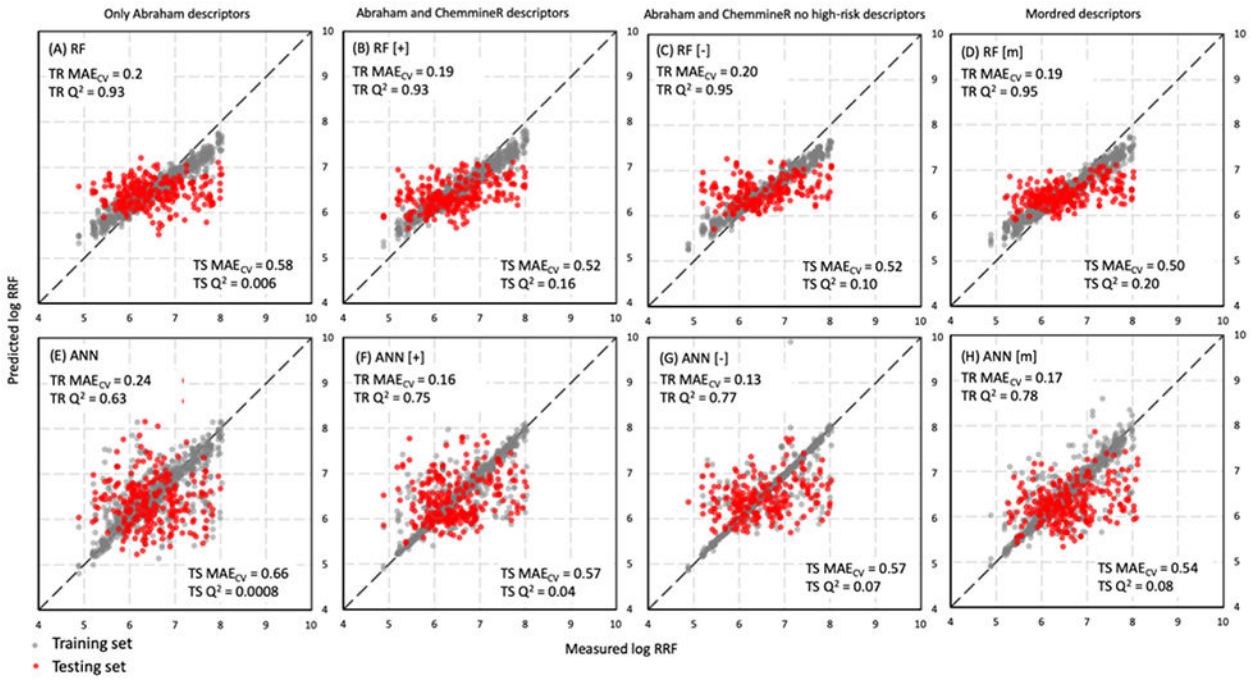

Figure 6:

5-fold CV analysis for the LC-QTOF/MS ESI− dataset for scenario 1 (only Abraham descriptors; A and E), scenario 2 (Abraham and ChemmineR descriptors; B and F), scenario 3 (Abraham and ChemmineR without high-risk descriptors; C and G) and scenario 4 (Mordred descriptors; D and H). The top panels (A-D) present the calculations for the RF models and the bottom panels (E-H) present the calculations for the ANN models. The 5-fold CV was conducted by randomly dividing the dataset into training and testing sets following an 80/20 split, building the model using the training set and testing it using the testing set. The process was repeated five times for each model and the models were assessed based on the CV mean absolute error (MAE_{CV}) of the predictions and the CV coefficient of determination (Q^2).

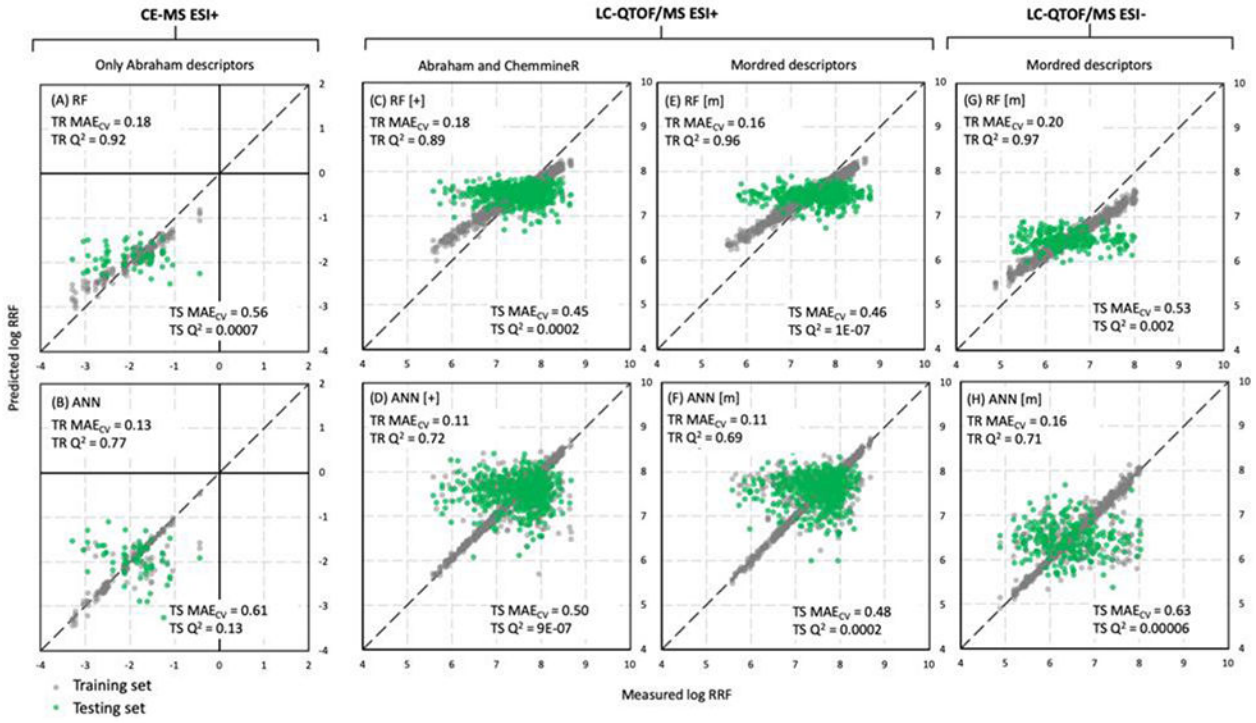

Figure 7:

Y-randomization for the best performing models for both the CE-MS ESI+ and the LC-QTOF/MS ESI+ and ESI− datasets. A and B: RF and ANN for the CE-MS ESI+ dataset using the Abraham descriptors. C and D: RF [+] and ANN [+] for the LC-QTOF/MS ESI+ dataset using the Abraham and ChemmineR descriptors. E and F: RF [m] and ANN [m] for the LC-QTOF/MS ESI+ dataset using the Mordred descriptors. G and H: RF [m] and ANN [m] for the LC-QTOF/MS ESI− dataset using the Mordred descriptors. For the y-randomization, we kept the X variable as it was and we randomized the y-variable. We divided the dataset into training and testing set following an 80/20 split, building the model using the training set and testing it using the testing set. The process was repeated five times for each model, as for the 5-fold CV, and the models were assessed based on the CV mean absolute error (MAE_{CV}) of the predictions and the CV coefficient of determination (Q^2).