SCIENTIFIC REPORTS
nature research

OPEN

Quantification for non-targeted LC/MS screening without standard substances

Jaanus Liigand[1], Tingting Wang[2], Joshua Kellogg[3], Jørn Smedsgaard[2], Nadja Cech[3] & Anneli Kruve[1,4*]

Non-targeted and suspect analyses with liquid chromatography/electrospray/high-resolution mass spectrometry (LC/ESI/HRMS) are gaining importance as they enable identification of hundreds or even thousands of compounds in a single sample. Here, we present an approach to address the challenge to quantify compounds identified from LC/HRMS data without authentic standards. The approach uses random forest regression to predict the response of the compounds in ESI/HRMS with a mean error of 2.2 and 2.0 times for ESI positive and negative mode, respectively. We observe that the predicted responses can be transferred between different instruments via a regression approach. Furthermore, we applied the predicted responses to estimate the concentration of the compounds without the standard substances. The approach was validated by quantifying pesticides and mycotoxins in six different cereal samples. For applicability, the accuracy of the concentration prediction needs to be compatible with the effect (e.g. toxicology) predictions. We achieved the average quantification error of 5.4 times, which is well compatible with the accuracy of the toxicology predictions.

Liquid chromatography mass spectrometry (LC/MS) has become the most versatile analytical tool to discover and detect metabolites[1], pharmaceuticals and their transformation products[2], environmental contaminants[3], and food contaminants[4] with non-targeted[5] analysis. To better understand the chemical and mechanistic dynamics of a system, a quantitative approach is preferred, which requires two main elements: identification and quantification of each metabolite (determining the concentration of compounds within the dataset). Currently, accurate mass measurements from high-resolution mass spectrometry (HRMS), together with relevant data analysis[6,7], are increasingly able to assign putative structures to the detected features[8]. Quantifying, however, remains a primary challenge. For example, out of 114 100 compounds in the Human Metabolome Database, only ca. 21 000 have been detected and identified[9]. Currently, the ability to obtain quantitative information from LC/MS is almost exclusively limited by the availability of standard substances, as different compounds ionize to different extents in an electrospray (ESI) source. The response of the compounds in LC/MS is influenced by the properties of the compound, eluent composition, and instrument. Thus, quantifying all detected compounds with a targeted analysis is exceedingly difficult, as standard reference materials (to match retention time, mass fragmentation pattern, and provide a calibration curve) are not available for the majority of compounds.

Additionally, the results of most LC/MS analyses conducted in different laboratories can currently only be compared based on qualitative data, as the measurement conditions and instruments used vary strongly and quantitative data are not available[5]. The lack of facile quantification also represents an obstacle to longitudinal studies, as samples collected over a long period of time must be stored and analysed all together in the same laboratory with the same methods. This raises concerns about sample preservation and stability, and delays in information dissemination, especially in cases where fast interventions may be crucial.

Several groups have studied ionization efficiencies in electrospray sources to understand which factors contribute to the ionization process[10–15]. Several ionization efficiency prediction models have been proposed for small sets of compounds, mainly based on multilinear regression algorithms (see Table S1 in Supporting Information

[1]Institute of Chemistry, Faculty of Science and Technology, University of Tartu, Ravila 14A, 50411, Tartu, Estonia. [2]National Food Institute, Research Group for Analytical Food Chemistry, Technical University of Denmark, Kemitorvet Building 202, Kgs, Lyngby, DK-2800, Denmark. [3]Department of Chemistry & Biochemistry, University of North Carolina at Greensboro, Greensboro, North Carolina, 27412, United States. [4]Department of Environmental Science and Analytical Chemistry, Stockholm University, Svante Arrhenius väg 16, 106 91, Stockholm, Sweden. *email: anneli.kruve@su.se

for a detailed overview). In general, researchers have focussed on finding a suitable predictive model for a specific compound group (drugs, metabolites, steroids) in studies that normally cover a few tens of compounds measured on a single instrument, and most have been carried out in ESI positive mode with one eluent composition in infusion or flow injection mode. More recently, some groups have used LC separation with gradient elution[16]. The latter point is important to effectively model the influence of both compound properties and eluent properties. However, a general model that would fit a wide range of compounds, LC conditions, instruments, and would also allow quantitation is lacking.

Based on the promising results obtained for predicting ionization efficiencies for specific compound classes, we propose a general approach for quantifying compounds using their predicted ionization efficiency values. We demonstrate that this approach can be transferred between different eluents, LC setups, and instruments and can be used for the analysis of real samples. The approach incorporates both positive and negative ionization modes under more than 100 eluent compositions and covers over 450 compounds. We validated this approach by predicting the concentrations of 35 compounds, including pesticides and mycotoxins, in cereal samples. The quantification for the samples was conducted on an instrument that was not involved in the ionization efficiency model development. The results presented herein demonstrate that our approach for calculating ionization efficiencies has the potential for making non-targeted analysis (semi-)quantitative.

Experimental

Ionization efficiency data: training and test set. Previously measured ionization efficiency values were collected in positive[17–24] and negative[24–28] mode. The ionization efficiency values for both modes and all studied eluent compositions are presented in Table S2. We measured an additional 165 compounds with diverse chemical properties in both the main eluent composition (acetonitrile/0.1% formic acid (aq) 80/20) and ca. 1000 new analyte-eluent combinations. For ESI positive mode, a total of 3139 ionization efficiency values were measured and collected from our previous works[17–21]. These data belong to 353 unique compounds and 106 different eluent compositions (Table S3 in SI). For ESI negative mode, an additional 1286 ionization efficiency values were collected from our previous works[25,26], including 33 eluent compositions Table S4), and 101 unique compounds.

The compounds covered by the model include protonated and intrinsically charged compounds in positive mode and deprotonated compounds in the negative mode. In ESI positive mode both nitrogen and a significant amount of oxygen bases could be measured. In ESI negative mode compounds with significant acidic moieties could be detected as $[M\text{-}H]^-$, including amino acids, benzoic acids and derivatives, phenols, amines, heterocycles, guanidines, and diazines (see Table S5 and S6). As we focus on $[M]^+$, $[M+H]^+$ and $[M\text{-}H]^-$, some of the compound groups are better represented then some of the other groups. For example, the lipids and sugars form primarily sodiated ions in ESI positive mode and are, therefore, not included in the scope of this approach.

From an application perspective, the training and test set compounds fall into a diverse array of categories, including drugs or drug-like compounds (e.g. terfenadine, ketoconazole, lidocaine), metabolites (e.g. acetylcholine, dopamine, thiamine), amino acids, small organic precursors (amines primarily), lipids (e.g. myristic acid, progesterone, glyceryl tributyrate) as well as industrial dyes (e.g. sudan II, sudan IV). The chemical space covered by these compounds was compared to the chemical databases Drugbank, NORMAN, and the Human Metabolome Database (HMDB), and is presented at Figs. S1 and S2.

Validation set. The proposed method was validated on a set of 35 pesticides and mycotoxins (Table S7), 28 of which had not been included in the training or test set. The chemical space covered by the validation compounds can be seen from Fig. S3. The compounds were measured at 10 concentration levels in solvent (acetonitrile), oat, barley, rye, wheat, rice, and maize with Agilent 6495 triple quadrupole instrument. The concentrations of the compounds ranged over 5 orders of magnitude from 3.6 nM to 0.35 mM. Altogether, 2233 data points (pesticide, matrix, and concentration combinations) were measured and corresponding concentrations were predicted. The validation set was used to evaluate the applicability of the ionization efficiency predictions for compounds not included in the training or test set. All measurements were also done under gradient separation and on an instrument that was not the primary instrument for ionization efficiency measurements.

Ionization efficiency measurements. For the evaluation of log*IE* values, the responses of $[M+H]^+$ and $[M\text{-}H]^-$ were recorded in an MS full scan mode corresponding to the polarity mode. In case in-source fragmentation of the compounds occurred, the intensities of the fragment ions peaks were summed with the intensity of the molecular ion peak. Six dilutions of the stock solutions were made (1-, 1.25-, 1.67-, 2-, 2.5-, and 5 times) with the corresponding eluent by the autosampler and delivered to MS in flow injection mode. The injection volume was 10 μL, and the eluent flow rate was 0.2 mL/min. Concentrations in the injected solutions ranged from 10^{-4} M to 10^{-9} M and measurements were conducted in the linear range. Most of the measurements of log*IE* values were carried out using an Agilent XCT ion trap mass spectrometer. Additionally, six mass spectrometers from five vendors were used in three labs around the world (Tartu/Estonia, Lyon/France, and Beerse/Belgium, see Table S8). On all instruments, the default MS and ESI parameters were used. In the case of the ion trap instrument, only the optimized Target Mass (TM) parameter was used[29].

The absolute ionization efficiency values vary significantly depending on the ionization source geometry, ion optics, day, cleanliness of the ionization source, etc. Therefore, we measured the relative ionization efficiency (*RIE*) of a compound M_1 relative to anchor compound (M_2) according to the following equation[26]:

$$RIE(M_1/M_2) = \frac{slope(M_1) \cdot IC(M_1)}{slope(M_2) \cdot IC(M_2)} \quad (1)$$

where the slope of the signal versus concentration is estimated via linear regression in the linear range of the signal-concentration plot and the *IC* is the sum of relative abundances of isotopologues where highest abundance is taken equal to 100. All measurements were made relative to tetraethylammonium (positive mode) and benzoic acid (negative mode). To make the data easier to present and analyze, the logarithmic scale (log*IE*) was used. The scale in negative mode was anchored to the log*IE* of benzoic acid, taken as 0 in the 0.1% ammonia/acetonitrile 20/80 mixture and the scale in positive mode was anchored to log*IE* of tetraethylammonium, taken as 3.95 in the 0.1% formic acid/acetonitrile 20/80. The log$IE = 0$ corresponds to methyl benzoate in the 0.1% formic acid/acetonitrile 20/80. The log*IE* values for each compound were obtained based on the *RIE* and log*IE* value of the anchor compound.

$$\log IE_{M_1} = \log RIE(M_1/M_2) + \log IE_{\text{anchor}} \tag{2}$$

To minimize the influence of possible differences in conditions when measuring M_1 and anchor compound, two steps were taken: (1) each compound was measured on at least three different runs (on three different days) and the results were averaged, and (2) anchor compound was measured in the beginning and the end of each run on each day. To anchor the scales of other eluent compositions, the MS signal intensities of anchor compound in all eluent compositions were measured in the same day and the log*IE* value of anchor compound in an eluent composition b was calculated using Eq:

$$\log IE_{\text{Eb}} = \log\left(IE_{\text{Ea}} \cdot \frac{Signal_{\text{Eb}}}{c_{\text{Eb}}} \cdot \frac{c_{\text{Ea}}}{Signal_{\text{Ea}}}\right) \tag{3}$$

where the $Signal_{\text{Eb}}$ and $Signal_{\text{Ea}}$ are the signal intensities in eluent composition b and a and c_{Ea} and c_{Eb} are the corresponding concentrations of benzoic acid in the respective eluent compositions.

To develop the model, log*IE* values measured on different instruments were transformed to unified log*IE* values. The unification was performed using the intersection of compounds measured on two instruments in the eluent composition acetonitrile/0.1% formic acid (aq) 80/20[22,23]. To unify the data, linear regression between log*IE* values measured on two instruments was performed (see SI).

Eluent effect of ionization efficiency. PCA analysis of the physicochemical properties (Fig. S4) of the training and test set compounds (n = 353) was conducted in order to incorporate different eluent compositions to the ionization efficiency predictive model without measuring all of the compounds in all eluents. Next, random sets of 40 compounds were sampled, and the set representing the widest chemical space was chosen for studying the solvent effects. As a result, 20 additional eluent compositions were studied (Table S3), yielding 106 different solvent compositions together with the previously measured eluent conditions. Both acetonitrile and methanol were studied as an organic modifier and the organic modifier percentages studied were 0%, 20%, 50%, 80% and 100%. Water phase additives included formic acid, trifluoroacetic acid, ammonia, ammonium acetate, ammonium formate, ammonium bicarbonate, and ammonium fluoride (pH = 2.0–10.3).

Model development. *Data preprocessing.* For model development, PaDEL descriptors[30] (1444 descriptors) were calculated using ChemDes calculator[31] for every compound. As PaDEL descriptor calculation of some descriptors fails for some compounds, the descriptors with NA (not available values) were removed from the dataset. Next, all the descriptors with the same value for >95% of compounds were eliminated from the dataset. As the third cleaning step, the pairwise correlation of descriptors was considered. If the R^2 was higher than 0.8 the former descriptor was removed from the dataset. After data pre-processing for ESI positive mode 1086 descriptors were left in the dataset and for ESI negative mode 822 descriptors were left in the dataset. Additionally, five empirical eluent descriptors (viscosity[32], surface tension[33], polarity index[34], pH, NH_4 content (yes/no)) were added to the dataset.

Algorithm selection and model development. Different machine learning algorithms (multilinear regression, Ridge regression, support vector machine regression, artificial neural networks and random forest regression) were tested for model development. For each algorithm, a suitable R package (glmnet[35], e1071[36], h2o[37], RRF[38]) was used and the parameters were optimized.

Regularized random forest regression algorithm[38] from library RRF in R yielded the best prediction accuracy. For ESI negative and positive mode, individual models were developed. For model development, the order of the dataset was randomized and split into two sets. 80% of observations were used for developing the model and 20% of observations were used as a validation set. The number of trees used in the random forest was optimized, with the optimal number was 100 decision trees. The regularization isotherm selected 450 significant descriptors (Table S9) in ESI positive mode and 145 significant descriptors (Table S10) in ESI negative mode.

The code used for training the model is available from Code S1.

Concentration from predicted ionization efficiency. As the model output is in universal ionization efficiency values and not instrumentation specific, a set of compounds with known concentrations is used to transform the universal predicted values to instrumentation specific values.

For transforming the predicted ionization efficiency values a set of compounds with known concentration either spiked to the matrix or as a standard solution was measured in dynamic range with the same method as compounds of interest. Sets of 6, 10, 15, and 31 compounds were tested; see discussion below. From the analysis results logarithmic response factors (*RF*) were calculated:

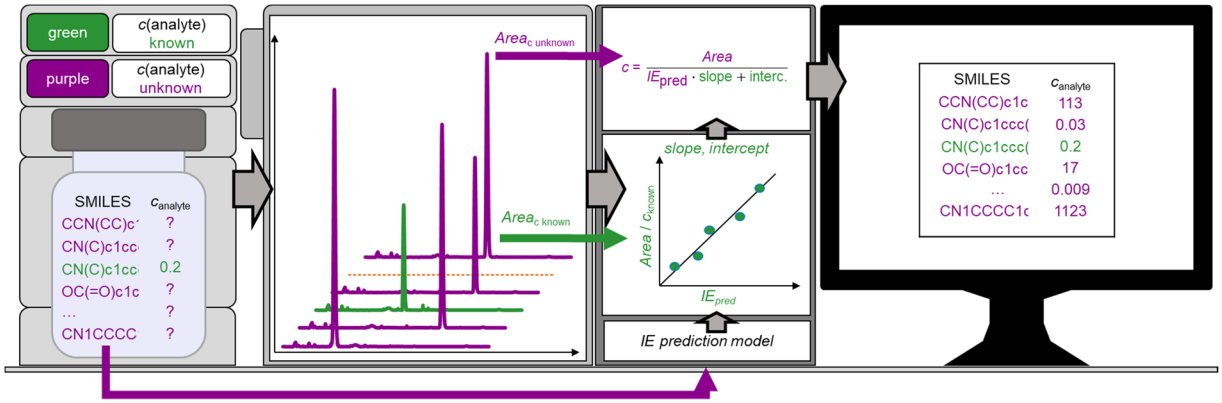


Figure 1. Flow chart of the developed approach to apply ionization efficiency prediction to estimate concentration. Purple is used for compounds of interest and green is used for compounds with known concentration; the latter are used to account for instrument-specific effects in the prediction model.

$$\log RF = \frac{Signal}{concentration} \tag{4}$$

and correlated with the predicted ionization efficiency values $\log IE_{\text{pred}}$:

$$\log RF_{\text{pred}} = slope \cdot \log IE_{\text{pred}} + intercept \tag{5}$$

where the signal is the MS1 peak area corrected with isotope distribution. Molar concentrations were used for the calculation of the response factor. Logarithmic response factors were correlated with predicted universal ionization efficiency values to obtain parameters necessary for transforming the predicted ionization efficiency values. The process of concentration prediction is visualized in Fig. 1. For studied pesticides and mycotoxins in cereals, the obtained parameters were used to transform predicted ionization efficiency values to logarithmic response factors. Based on the MS1 peak areas of compounds of interest and predicted response factors it is possible to predict the concentration. Slope and intercept values in Eq. 5 were calculated based on the coefficients of the linear regression curve between $\log RF$ and $\log IE_{\text{pred}}$ values in the transformation set. In order to validate the obtained results, the prediction errors between real concentration and predicted concentration were calculated.

Measuring the prediction accuracy. We express the accuracy of ionization efficiency prediction as a prediction error:

$$\text{prediction error} = \max\left\{\frac{predicted\ IE}{measured\ IE}, \frac{measured\ IE}{predicted\ IE}\right\} \tag{6}$$

The lowest possible prediction error is 1.0 times and values close to 1.0 times are desirable. In order to illustrate the general prediction accuracy, we use the average error. The accuracy of concentration predictions was estimated with a similar comparison.

Sample preparation for cereal samples. All of the cereal samples were prepared according to the QuEChERS extraction method, described elsewhere[39]. In short, the blank samples (pesticides free) were obtained from proficiency test material for the six European Union Proficiency Test EUPTs: EU-PT-CF8 (wheat), C3 (oat), CF10 (rye) and C6 (barley), CF9 (maize), SRM6 (rice). Two grams of homogenized cereal samples were soaked with 10 mL acidified Milli-Q water containing 0.2% formic acid. Then, the sample was extracted with 10 mL of acetonitrile. Thereafter, 4 g of magnesium sulfate and 1 g of sodium chloride were added, and the tube was shaken for 1 min followed by centrifugation. The organic upper layer (2 mL) was removed and shaken with 0.1 g of Bondesil-C18 and 0.3 g of magnesium sulphate for 2 min followed by centrifugation. Then 1.5 mL of purified extract was removed into a vial with insert and spiked with different concentration of tested compounds prior to injection on the LC/MS system.

LC/MS analysis of cereal samples. Samples were analyzed on an Agilent 1290 ultrahigh performance liquid chromatograph (Agilent Technologies, CA, U.S.) coupled to an Agilent 6495 triple quadrupole instrument (QQQ) at the University of Tartu (UT). Samples were injected onto an Agilent Zorbax RRHD SB-C18 reversed-phase column (1.8 μm, 2.1 × 50 mm). The mobile phase consisted of (A) water containing 0.1% formic acid and (B) acetonitrile. The analysis was done using a gradient elution at a flow rate of 0.3 mL/min at 30 °C. The gradient was from 5% to 100% B in 7 min, then maintained at 100% for 2 min and returned back to 5% B in 2 min, and maintaining starting conditions at 5% B for 2 min equilibration with 5% B to yield a total runtime of 13 min.

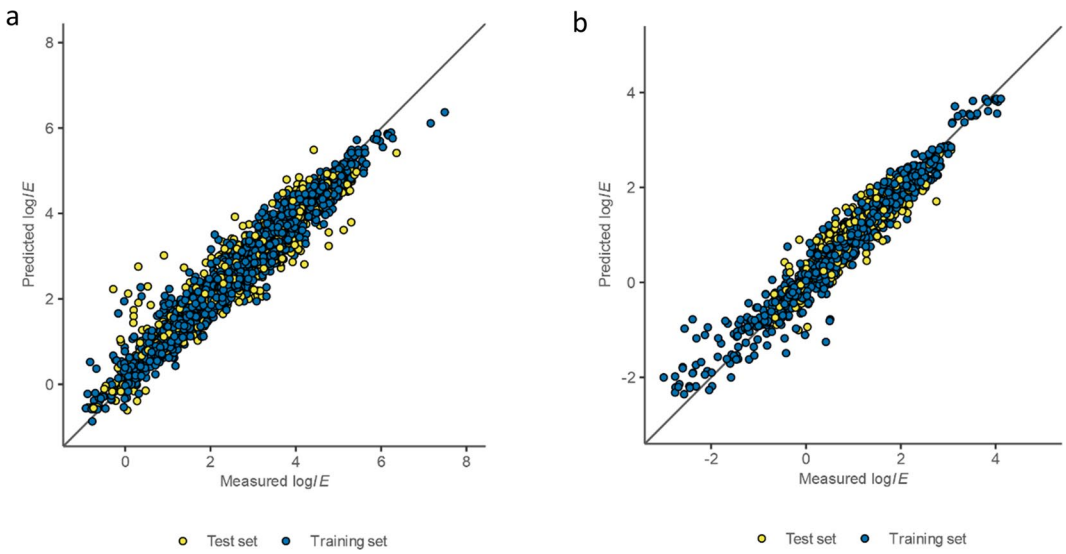


Figure 2. Performance of ionization efficiency prediction models. Black line denotes ideal fit. (**a**) ESI positive mode with 3139 datapoints. (**b**) ESI negative mode with 1286 datapoints.

An Agilent's Jet Stream electrospray ionization (ESI) interface was used in positive ion mode (ESI+) with the following settings: capillary voltage 3 kV, nebulizer pressure 20 psi, sheath gas flow rate 11 L/min, sheath gas temperature 350 °C, dry gas temperature 250 °C, and dry gas flow rate 14 L/min. Spectra were collected from m/z 100 to 1100 Da. The injection volume was 1 μL. Peaks were manually integrated, and peak areas were used for all calculations.

Results

Predicting ionization efficiencies. In order to quantify the compounds based on the ionization efficiency values we (1) measured *IE* values for a wide set of compounds, (2) used the measured *IE* values as well as compound and eluent descriptors for developing the model that would allow predicting ionization efficiencies, and (3) validated the approach by using the predicted *IE* to quantify a set of compounds in cereal samples. All of the *IE* values have been measured as relative values; the *IE* of methyl benzoate and benzoic acid have been arbitrarily taken as $IE = 1$ ($\log IE = 0$) in ESI positive and negative mode, respectively.

Based on these ionization efficiencies, predictive models for positive and negative mode were developed. The overall root mean squared prediction error was 2.2 times (training set 1.9 and test set 3.0 times) (Fig. 2a and Table S11). This means that if the ionization efficiency of compound A is predicted to be 100 times higher than the ionization efficiency of the methyl benzoate the actual ionization efficiency would be 45 to 220 higher than that of methyl benzoate ($\log IE = 2.00 \pm 0.34$).

In negative mode, the best performing model was obtained also with random forest regression with 145 significant descriptors and 100 regression trees. The regression model explained 93% of the variation in ionization efficiency values. The overall root mean squared error was 2.0 times (training set 2.0 and test set 2.3 times, Fig. 2b).

Upon closer examination of the ionization efficiency prediction model in ESI positive mode, it was observed that the model performed universally well for different organic modifier percentages (Fig. S5). The lowest prediction error, 1.4 times, was observed for eluents containing 20% of organic modifier and the highest prediction error, 1.9 times, for eluent containing 90% organic modifier. This is expected, as eluents containing 20% of organic modifier had the highest number of data points, which improved prediction accuracy. Additionally, the model was well-performing for both methanol as well as for acetonitrile containing eluent compositions (Figs. S5 and S6). Based on the pH of the eluent, basic conditions had the highest prediction error; 2.5 times and 3.7 times for the training and test set, respectively.

Similar trends were observed for ESI negative mode; the prediction accuracy for the pure organic modifier was the lowest (prediction error of 4.1 times, Figs. S7 and S8). Regarding the pH, no significant differences in the prediction accuracy were observed (Fig. S7).

The most influential parameters (Table S12) were a number of structural parameters and also solvent parameters. PaDEL descriptors used in the Random Forest regression are 2D structural parameters and are, therefore, not directly linked to classical physicochemical parameters; however, some of the most influential parameters can be interpreted. From structure-related parameters, number of hydrogen atoms and number of nitrogen atoms was important in the model in ESI positive mode. The number of hydrogen atoms is associated with the size and hydrophobicity of the compounds while the number of nitrogen atoms is associated with the basicity of the compound. Also, the mobile phase parameters pH, viscosity, and presence of ammonium ions in the mobile phase were significant. Previous studies have shown that these factors can influence the response of compounds by orders of magnitude[18–20]. Here, the model confirms the previous empirical findings.

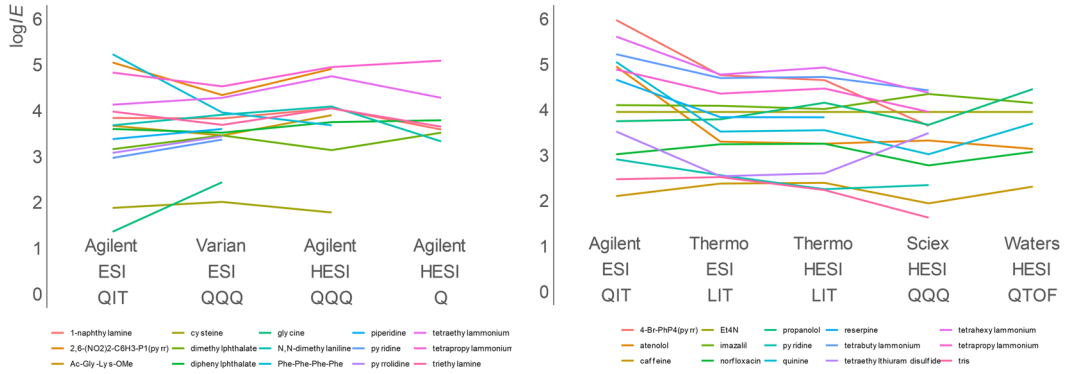


Figure 3. Measured ionization efficiencies on different instruments in ESI positive mode based on two subsets of druglike compounds measured on different instruments in the same eluent composition (acetonitrile/0.1% formic acid(aq) 80/20). Full data are shown in Table S2. ESI denotes conventional pneumatically assisted electrospray ionization source, HESI heated electrospray ionization source, QIT quadrupole ion trap, QQQ triple quadrupole, LIT linear iontrap, QTOF quadrupole time-of flight. Left and right correspond to two datasets.

Compatibility between different instruments. It is known that a compound's response varies from one LC/MS instrument to another; however, we have previously shown that the relative log*IE* values measured on different instruments are in good correlation to each other (Figs. 3 and S9)[22,23]. The ionization efficiency model was primarily developed on an ion trap instrument; however, for this study, we expanded the types of LC/MS instruments used for measuring ionization efficiency values. Altogether seven instruments and 9 different instrument-ionization source combinations (Table S8) and different mass analyser types from labs around the world have been included in studying the ionization efficiency. On all instruments, generic default settings have been used. It is known that some instruments (e.g. QQQ) may have lower mass cut-offs. Here, these effects are incorporated in the agreement of ionization efficiency values.

The comparison of the ionization efficiency values for compounds measured on a number of instruments reveals good intra-instrumental consistency (Figs. 3 and S9). The compounds being highest responders on one instrument are also generally top performers on the other instruments. Still, some instruments compress the ionization efficiency values more than other instruments. However, these effects could not be associated with a known lower mass cut-off of mass analysers or ionization source type (ESI vs HESI). Therefore, a more complicated relationship between the source design and ion optics design is likely to be important.

At the same time, it is not reasonable to build time and data-intensive prediction algorithms on each instrument. Therefore, it is necessary to have a robust approach to transfer the predicted values between different instruments. In order to achieve this, we use a linear correlation observed between the ionization efficiency values measured on different instruments (see SI for more details).

The feasibility of using linear correlation for transferring the predicted ionization efficiency values can be estimated by comparing the prediction error of ionization efficiency values for different instruments. We observed a good consistency between the prediction errors from instrument to instrument (Fig. S10). The highest average prediction error was 2.8 times (Waters Synapt G.2 Z-spray) and the lowest was 1.5 times (Thermo LTQ ESI). Both of these values are close to the repeatability of the log*IE* values. Therefore, it can be concluded that ionization efficiency values can be transferred between instruments with reasonable accuracy.

Additionally, we developed a model to predict ionization efficiency values with PaDEL descriptors and random forest regression based on data from a single instrument. For this, we chose data from an Agilent XCT instrument, which had been used to measure the largest number of ionization efficiency values (1816 data points). The mean prediction error for a single instrument-based model (ESI + mode) was 1.8 and 3.0 times for training and test set, respectively. The achieved prediction error values for a single instrument-based model were not significantly different from the generic model incorporating data from seven different instruments. Therefore, the developed model is robust for variations caused by different instruments and can be applied to predict ionization efficiency values measured on different mass spectrometers.

Validation: ionization efficiency and quantification of pesticides and mycotoxins in cereals. The proposed was validated based on 35 pesticides and mycotoxins (Table S7) measured under gradient separation on a triple quadrupole instrument. The list included 28 compounds not included in the training or test set. The log*IE* values were measured from the analysis of standard solutions of the 35 validation compounds and the values range from 1.60 to 4.11 with the average of 3.10 and median 3.29. The predicted log*IE* values take into account the eluent composition at retention time. The log*IE* prediction with the random forest model showed reasonable accuracy (RMSE 0.64, Fig. S11).

The predicted ionization efficiency was further used to predict the concentrations of the compounds detected assuming that the structure is known. For this, the compounds were spiked into blank oat, barley, rye, wheat, rice, and maize and analysed. For each compound, the ionization efficiency was predicted in ESI positive mode. To transform the predicted ionization efficiencies to instrument-specific response factors, a set of 31 compounds (Table S7) was used. Thereafter, the instrument-specific response factors were used to convert LC/MS signals into concentration (Eq.).

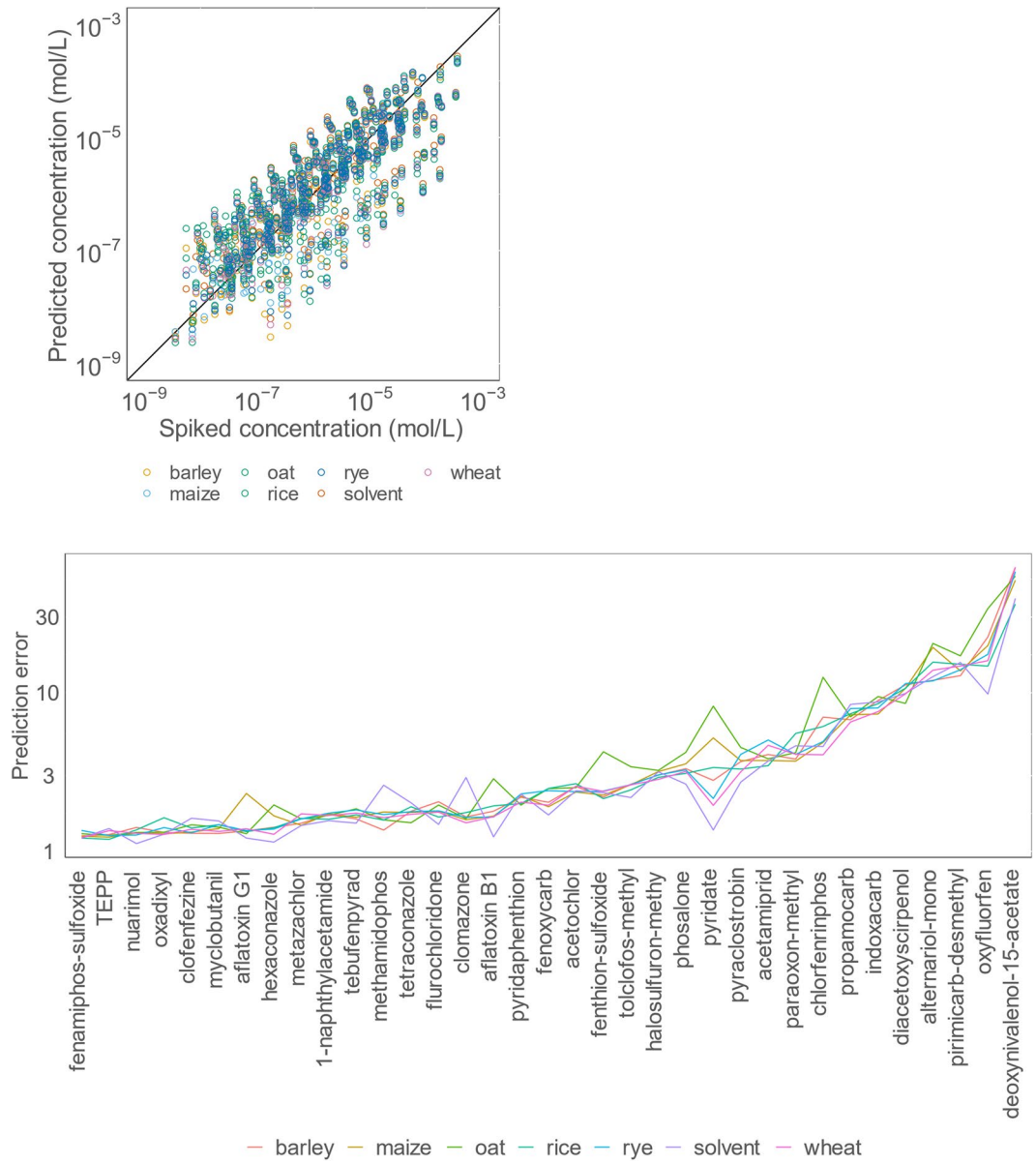


Figure 4. Performance of concentration prediction in the example of pesticides in cereals. above: concentration prediction of pesticides in cereal samples. below: prediction error of pesticide concentration in cereal samples, y-axis in logarithmic scale.

$$c = \frac{Signal}{10^{\log RF_{pred}}} \tag{7}$$

On average, the concentrations were predicted with the prediction error of 5.4 times (Fig. 4a, compound separated graphs on Fig. S12, Table S13). This means that if the pesticide concentration is estimated to be 1 ppm it would actually lie between 0.2 and 6 ppm. Compared to the conventional approach of assuming the equal response to all compounds detected (average prediction error of 526 times, Table S14), the developed approach improved prediction accuracy around 10 times and significantly reduces the width of the confidence interval. Consequently, more effective decision making can be made based on the predicted concentrations.

The lowest error observed was 1.2 times for TEPP in rice matrix, while the largest error, 62 times was observed for deoxynivalenol-15-acetate in barley matrix. For 88% of the compounds, the prediction error was lower than 10 times, for 76% of the compounds it was lower than 5 times, and for 48% of the compounds, it was lower than 2 times. The three compounds with the highest prediction error were deoxynivalenol-15-acetate (52 times), oxyfluorfen (19 times), and pirimicarb-desmethyl (15 times). The three best-performing compounds are fenamiphos-sulfoxide (1.2 times), TEPP (1.3 times), and nuarimol (1.3 times). These prediction errors are especially promising for mycotoxins for which the standard substances are either rarely available or very expensive. The large errors could not be associated with specific compound properties.

Additionally, we tested sets of 15, 10 and 6 compounds for transforming the predicted ionization efficiency values to instrument-specific response factors. For this random sets of 15, 10 or 6 compounds were sampled from the 31 compounds. On average, sets of 15, 10, and 6 compounds resulted in concentration prediction errors of 5.0, 5.0, and 5.0 times, respectively, which were not significantly different from the average prediction error observed for 31 compounds. Therefore, comparable prediction accuracies can be achieved with a smaller set of compounds. However, it needs to be considered that the compounds used should possess a wide range of chemical properties as well as a wide range of ionization efficiency values to allow for the establishment of a reasonable regression approach. Additionally, uniform elution distribution during the chromatographic run is beneficial.

Different matrices. Any model that aims at providing quantitative information needs to be applicable in a variety of matrices in order to be truly useful to researchers. The importance of matrices with LC/MS is even more relevant than for other analytical techniques, due to the possibility of significant matrix effects in the ESI source. A matrix effect is the suppression of ionization of a compound due to co-eluting compound(s). Previously, it has been qualitatively observed that the matrix effect and ionization efficiencies are influenced by the physicochemical properties of the compound[40]. Therefore, while applying the ionization efficiency predictions for concentration estimations, it was assumed that a model that considers ionization efficiencies also helps to account for matrix effect. This assumption was based on the fact that a small set of compounds with known concentrations were spiked into every sample; this helped to account for the differences arising from the instrument and also for matrix effects (Eq. 7).

Regarding different matrices, the prediction accuracy for all cereals was very similar. The lowest prediction error was observed for wheat and rice, 4.8 times, and highest for oat, 6.3 times (Table S15). This is expected, as oat samples possess a high content of polar lipids and free fatty acids[41] which possess high surface affinity and are, therefore, expected to cause ionization suppression[42]. Also, the mean prediction error for solvent (average 4.6 times) and all studied cereals (5.5 times) was comparable.

Moreover, the matrix effect is expected to vary strongly from sample to sample even for the same food commodity. The compounds that performed worst in one matrix performed also performed poorly in other matrices, and the best performers were among the top for all matrices (Fig. 4b). The consistency of the prediction error form one matrix to another indicates the insignificance of matrix effect and strongly indicates that the developed approach using ionization efficiency predictions, together with the transformation, helped to account for matrix effects.

Application area. The application range of all models depends on the data that have been used to train the model. The compounds covered by the model include protonated $[M+H]^+$ and intrinsically charged compounds $[M]^+$ in positive mode and deprotonated $[M-H]^-$compounds in the negative mode. This defines the limitations of the application range of the model. For examples, compounds forming primarily sodium or ammonium adducts and no protonated ions are not within the scope of this approach.

We were additionally interested in comparing the compounds used in this study to compounds detectable with LC/MS based non-targeted screening. For this purpose, we compared the properties of the compounds used in this study with some of the most common databases; namely, NORMAN database[43], HMDB[9], and DrugBank[44]. NORMAN database is a database compiled by NORMAN network members and contains compounds that are important for various application areas from solvents to industrial chemicals, personal care products to metabolites. HMDB contains information about metabolites with different origins and DrugBank contains approved small molecule drugs, approved biologics (proteins, peptides, vaccines, and allergenics), nutraceuticals and experimental (discovery-phase) drugs. It is important to consider that LC/MS is unable to analyse all of these compounds due to limitations in retention and ionization. For HMDB and NORMAN it was possible to extract only the compounds that had been analysed with LC/MS. The compounds included in this study were compared with the databases based on the principal component analysis of PaDEL descriptors. It was observed that compounds included in this study covered a very large part on the chemical space of interest in HMDB and NOMRAN (Figs. S2 and S3). The coverage was somewhat less for DrugBank. This was expected, as the DrugBank includes also large biomolecules out of the scope of the model. Therefore, it is important to keep in mind the limitations of the initial data used in the modelling while applying the model and we hope to widen the model application are further in the future by incorporating (1) compounds forming adducts and (2) larger compounds, such as peptides.

Discussion

For the case study presented here, quantifying a variety of pesticides and mycotoxins in cereal matrices, we have shown that ionization efficiencies can be used to improve the accuracy of concentration predictions for small molecules that lack standard substances. In general, to obtain a reliable model, it is suggested to include several compounds that have been analysed together with the sample and quantified. Such compounds could be compounds confirmed with the aid of standard substances or compounds from quality control samples. This makes full scan LC/HRMS extremely appealing, as a combined targeted and non-targeted analysis method can be used.

Additionally, for risk assessment of contaminants in food and environmental samples, an estimated concentration of a compound is required to evaluate the potential risk of the exposure. In most cases, the error associated with other parts of the risk assessment (like intake and toxicology) is even larger than the error in concentration prediction[45]. Therefore, the ionization efficiency based quantification approach has high potential to complement non-targeted analysis and aid decision making based on the analysis results. The tool is made available at app.quantem.co. The approach has been developed on several instruments and validated on the example of five complicated matrices. Therefore, the quantification accuracy also represents the effects arising from mass discrimination and matrix effect. Validation for other applications is though, desirable.

Data availability

All data used for model development and validation are available as supporting information.

Received: 20 June 2019; Accepted: 16 March 2020;
Published online: 02 April 2020

References

1. Cajka, T. & Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. *Anal. Chem.* **88**, 524–545 (2016).
2. Xu, R. N., Fan, L., Rieser, M. J. & El-Shourbagy, T. A. Recent advances in high-throughput quantitative bioanalysis by LC–MS/MS. *J. Pharm. Biomed. Anal.* **44**, 342–355 (2007).
3. Alder, L., Greulich, K., Kempe, G. & Vieth, B. Residue analysis of 500 high priority pesticides: Better by GC–MS or LC–MS/MS? *Mass Spectrom. Rev.* **25**, 838–865 (2006).
4. Malik, A. K., Blasco, C. & Picó, Y. Liquid chromatography–mass spectrometry in food safety. *J. Chromatogr. A* **1217**, 4018–4040 (2010).
5. Schymanski, E. L. *et al*. Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal. Bioanal. Chem.* **407**, 6237–6255 (2015).
6. Allen, F., Pon, A., Wilson, M., Greiner, R. & Wishart, D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.* **42**, W94–W99 (2014).
7. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation. *J. Cheminformatics* **8**, 3 (2016).
8. Djoumbou-Feunang, Y. *et al*. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminformatics* **11**, 2 (2019).
9. Wishart, D. S. *et al*. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
10. Chalcraft, K. R., 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.* **81**, 2506–2515 (2009).
11. Cech, N. B. & Enke, C. G. Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides. *Anal. Chem.* **72**, 2717–2723 (2000).
12. Wu, L. *et al*. Quantitative structure–ion intensity relationship strategy to the prediction of absolute levels without authentic standards. *Anal. Chim. Acta* **794**, 67–75 (2013).
13. Huffman, B. A., Poltash, M. L. & Hughey, C. A. Effect of Polar Protic and Polar Aprotic Solvents on Negative-Ion Electrospray Ionization and Chromatographic Separation of Small Acidic Molecules. *Anal. Chem.* **84**, 9942–9950 (2012).
14. Henriksen, T., Juhler, R. K., Svensmark, B. & Cech, N. B. The relative influences of acidity and polarity on responsiveness of small organic molecules to analysis with negative ion electrospray ionization mass spectrometry (ESI-MS). *J. Am. Soc. Mass Spectrom.* **16**, 446–455 (2005).
15. Ehrmann, B. M., Henriksen, T. & Cech, N. B. Relative importance of basicity in the gas phase and in solution for determining selectivity in electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **19**, 719–728 (2008).
16. Hermans, J., Ongay, S., Markov, V. & Bischoff, R. Physicochemical Parameters Affecting the Electrospray Ionization Efficiency of Amino Acids after Acylation. *Anal. Chem.* **89**, 9159–9166 (2017).
17. Oss, M., Kruve, A., Herodes, K. & Leito, I. Electrospray Ionization Efficiency Scale of Organic Compounds. *Anal. Chem.* **82**, 2865–2872 (2010).
18. Liigand, J., Kruve, A., Leito, I., Girod, M. & Antoine, R. Effect of Mobile Phase on Electrospray Ionization Efficiency. *J. Am. Soc. Mass Spectrom.* **25**, 1853–1861 (2014).
19. Liigand, J., Laaniste, A. & Kruve, A. pH Effects on Electrospray Ionization Efficiency. *J. Am. Soc. Mass Spectrom.* **28**, 461–469 (2017).
20. Ojakivi, M., Liigand, J. & Kruve, A. Modifying the Acidity of Charged Droplets. *ChemistrySelect* **3**, 335–338 (2018).
21. Kruve, A. & Kaupmees, K. Adduct Formation in ESI/MS by Mobile Phase Additives. *J. Am. Soc. Mass Spectrom.* **28**, 887–894 (2017).
22. Liigand, J., Vries, Rde & Cuyckens, F. Optimization of flow splitting and make-up flow conditions in liquid chromatography/electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **33**, 314–322 (2019).
23. Liigand, J. *et al*. Transferability of the Electrospray Ionization Efficiency Scale between Different Instruments. *J. Am. Soc. Mass Spectrom.* **26**, 1923–1930 (2015).
24. Liigand, P. *et al*. Think Negative: Finding the Best Electrospray Ionization/MS Mode for Your Analyte. *Anal. Chem.* **89**, 5665–5668 (2017).
25. Kruve, A. & Kaupmees, K. Predicting ESI/MS Signal Change for Anions in Different Solvents. *Anal. Chem.* **89**, 5079–5086 (2017).
26. Kruve, A., Kaupmees, K., Liigand, J. & Leito, I. Negative Electrospray Ionization via Deprotonation: Predicting the Ionization Efficiency. *Anal. Chem.* **86**, 4822–4830 (2014).
27. Kruve, A. Influence of mobile phase, source parameters and source type on electrospray ionization efficiency in negative ion mode: Influence of mobile phase in ESI/MS. *J. Mass Spectrom.* **51**, 596–601 (2016).
28. Liigand, P., Liigand, J., Cuyckens, F., Vreeken, R. J. & Kruve, A. Ionisation efficiencies can be predicted in complicated biological matrices: A proof of concept. *Anal. Chim. Acta* **1032**, 68–74 (2018).
29. Kruve, A., Kaupmees, K., Liigand, J., Oss, M. & Leito, I. Sodium adduct formation efficiency in ESI source: Sodium adduct formation efficiency in ESI source. *J. Mass Spectrom.* **48**, 695–702 (2013).
30. Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **32**, 1466–1474 (2011).
31. Dong, J. *et al*. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *J. Cheminformatics* **7** (2015).
32. Snyder, L. R., Kirkland, J. J. & Dolan, J. W. *Introduction to Modern Liquid Chromatography*. (John Wiley & Sons, Inc. (2009).
33. Rudakov, O. B., Belyaev, D. S., Khorokhordina, E. A. & Podolina, E. A. Surface tension of binary mobile phases for liquid chromatography. *Russ. J. Phys. Chem. A* **81**, 366–369 (2007).
34. Katz, E., Eksteen, R., Schoenmakers, P. & Miller, N. *Handbook of HPLC*. (M. Dekker (1998).
35. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
36. Meyer, D. Support Vector Machines. https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf.
37. *h2oai/h2o-3*. (H2O.ai (2019).
38. Deng, H. & Runger, G. Feature Selection via Regularized Trees. *ArXiv12011587 Cs Stat* (2012).
39. Dzuman, Z., Zachariasova, M., Veprikova, Z., Godula, M. & Hajslova, J. Multi-analyte high performance liquid chromatography coupled to high resolution tandem mass spectrometry method for control of pesticide residues, mycotoxins, and pyrrolizidine alkaloids. *Anal. Chim. Acta* **863**, 29–40 (2015).

40. Trufelli, H., Palma, P., Famiglini, G. & Cappiello, A. An overview of matrix effects in liquid chromatography–mass spectrometry. *Mass Spectrom. Rev.* **30**, 491–509 (2011).
41. Krasilnikov, V. N., Batalova, G. A., Popov, V. S. & Sergeyeva, S. S. Fatty Acid Composition of Lipids in Naked Oat Grain of Domestic Varieties. *Russ. Agric. Sci.* **44**, 406–408 (2018).
42. Ismaiel, O. A., Halquist, M. S., Elmamly, M. Y., Shalaby, A. & Thomas Karnes, H. Monitoring phospholipids for assessment of ion enhancement and ion suppression in ESI and APCI LC/MS/MS for chlorpheniramine in human plasma and the importance of multiple source matrix effect evaluations. *J. Chromatogr. B* **875**, 333–343 (2008).
43. Stravs, M. A., Schymanski, E. L., Singer, H. P. & Hollender, J. Automatic recalibration and processing of tandem mass spectra using formula annotation: Recalibration and processing of MS/MS spectra. *J. Mass Spectrom.* **48**, 89–99 (2013).
44. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
45. Blacquière, T., Smagghe, G., van Gestel, C. A. M. & Mommaerts, V. Neonicotinoids in bees: a review on concentrations, side-effects and risk assessment. *Ecotoxicology* **21**, 973–992 (2012).

Acknowledgements

This work was financially supported by Smart specialization doctoral stipend, PRG300 from Estonian Research Council, Otto Mønsted Foundation, and Fødevareforlig 3. To all financing sources the authors are greatly indebted. We thank Stefan Schoder and Asko Laaniste for fruitful discussions. Open access funding provided by Stockholm University.

Author contributions

J.L. carried out most of the ionization efficiency measurements, all of the modelling, and wrote significant part of the manuscript. T.W. carried out all of the measurements for validation, including concentration prediction of pesticides. J.S. helped to plan and analyse data related to pesticide analysis. J.K. and N.C. helped in concept verification. A.K. designed the project, helped to plan the experiments and modelling and wrote significant part of the manuscript. All authors contributed significantly to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-62573-z.

Correspondence and requests for materials should be addressed to A.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Quantification for non-targeted LC/MS screening without standard substances

Jaanus Liigand[1], Tingting Wang[2], Joshua J. Kellogg[3], Jørn Smedsgaard[2], Nadja B. Cech[3], Anneli Kruve[1,4]

[1] Institute of Chemistry, Faculty of Science and Technology, University of Tartu, Ravila 14A, 50411, Tartu, Estonia

[2] National Food Institute, Research Group for Analytical Food Chemistry, Technical University of Denmark, Kemitorvet Building 202, Kgs. Lyngby, DK-2800, Denmark

[3] Department of Chemistry & Biochemistry, University of North Carolina at Greensboro, Greensboro, North Carolina 27412, United States

[4] Department of Environmental Science and Analytical Chemistry, Stockholm University, Svante Arrhenius väg 16, 106 91 Stockholm

Methods

Eluent parameters

The viscosity of organic modifier water binary mixture[1–3] was calculated using the general model:

$$viscosity\ (\ mPa \cdot s) = A \cdot (org\% \cdot 100)^2 + B \cdot org\% \cdot 100 + C \qquad \text{Eq. 1}$$

The surface tension of organic modifier water binary mixture[4,5] was calculated using the general model:

$$\begin{aligned} surface\ tension\ (\mathrm{mN/m}) \\ &= \sigma_1 + D \cdot \sigma_1 \cdot org\% + (E \cdot \sigma_2 - D \cdot \sigma_1 - \sigma_1) \cdot org\%^2 \\ &+ (\sigma_2 - D \cdot \sigma_2) \cdot org\%^3 \end{aligned} \qquad \text{Eq. 2}$$

Polarity index of organic modifier water binary mixture[6] was calculated using the general model:

$$polarity\ index = F \cdot org\% + G \cdot (1 - org\%) \qquad \text{Eq. 3}$$

NH_4^+ presence parameter was 1 if the additive included either ammonia or ammonium salt (ammonium acetate, ammonium formate, ammonium bicarbonate, ammonium fluoride), otherwise, it was kept 0.

Parameters used to calculate eluent descriptors.

name	A	B	C	D	E	F	G	σ_1	σ_2
acetonitrile	$-1.04 \cdot 10^{-4}$	$4.36 \cdot 10^{-3}$	$8.84 \cdot 10^{-1}$	-2.9	7.14	5.8	–	–	27.9
methanol	$-3.59 \cdot 10^{-4}$	$3.20 \cdot 10^{-2}$	$9.03 \cdot 10^{-1}$	-2.2	5.62	5.1	–	–	22.1
isopropanol	$-4.74 \cdot 10^{-4}$	$5.89 \cdot 10^{-2}$	$7.88 \cdot 10^{-1}$	-3.9	15.6	3.9	–	–	17
acetone	$-3.13 \cdot 10^{-4}$	$2.47 \cdot 10^{-2}$	$9.02 \cdot 10^{-1}$	-2.5	6.84	5.1	–	–	22.2
water	–	–	–	–	–	–	10	71.8	–

Data unification between different instruments

To develop the model, log*IE* values measured on different instruments were transformed to unified log*IE* values. The unification was performed using the intersection of compounds measured on two instruments in the eluent composition acetonitrile/0.1% formic acid (aq) 80/20. A calibration graph was constructed logIE (Instrumentn) vs log*IE* (Agilent XCT). The log*IE* values measured on different instrument than Agilent XCT ion trap were transformed as follows:

$$\log IE\ (Agilent\ XCT) = slope \cdot \log IE\ (Instrument_n) + intercept \qquad \text{Eq. 4}$$

Table S1 In literature available studies focusing on ionization efficiency investigation and modelling.

Ref.	Compound type	Number of unique compounds	Eluent	Mode	Model	Matrix
[7]	drug like	10	0.1 % formic acid/acetonitrile 60/40	+	no	solvent
[8]	steroids	30	100 % methanol; ammonia (pH = 10)/methanol 50/50	+/−	PLS	solvent
[9]	oligonucleotide	11	water/methanol 50/50 with ion-pairing reagents	−	PLS	solvent
[10]	metabolites	9	0.5 mM ammonium acetate/methanol 1/9	+	no	urine
[11]	natural products	8	0.1 % formic acid/acetonitrile	+	no	solvent
[12]	lipids	15	100 % methanol	+	no	solvent
[13]	drugs	170	1 mM ammonium acetate/acetonitrile 50/50	+	PCA, Projection Pursuit. Classification and Regression Trees. PLS and Stepwise MLR	solvent
[14,15]	oligopeptides	23	0.5 % acetic acid/methanol 50/50	+	no	solvent
[16]	metabolites	58	0.1 % formic acid/methanol 50/50	+	MLR	solvent; RBC lysates

Ref.	Compound type	Number of unique compounds	Eluent	Mode	Model	Matrix
[17]	lipids	8	5 mM ammonium acetate/acetonitrile	+	no	NIST human plasma; egg yolk; porcine liver
[18]	drugs	77	0.1 % formic acid/acetonitrile	+	MLR	solvent
[19]	drugs	40	0.1 % acetic/acetonitrile (0.1 % acetic)	+	no	solvent
[20]	druglike	19	0.5 % acetic acid in methanol	+	no	solvent
[21]	druglike	20	7 mM acetic acid /methanol 20/80; 7 mM ammonia/methanol 20/80	+/−	ANN	solvent
[22]	amino acids	4		+	no	no
[23]	natural products	25	10 mM ammonium formate/acetonitrile 60/40; 50/50; 40/60; 30/70; 20/80	− (formate adduct)	no	solvent, *S. habrochaites* leaves
[24]	drugs	110	water/acetonitrile 20/80	−	PLS	solvent
[25]	sartans	7	pH = 5.5/methanol 40/60	+	ANN	solvent
[26]	drugs	71	water/acetonitrile 50/50	+	no	solvent
[27]	phenols	35	100 % methanol; 100 % acetonitrile; methanol/water 50/50; acetonitrile/water 50/50	−	no	solvent
[28]	amino acids. derivates	127	acetonitrile (0.1 % formic acid)/0.1 % formic acid	+	MLR	solvent
[29]	druglike	49	100 % acetonitrile; 100 % acetone;	−	no	solvent

Ref.	Compound type	Number of unique compounds	Eluent	Mode	Model	Matrix
			100 % methanol; 100 % water			
[30]	natural products	13	5 mM ammonium acetate/ methanol (5 mM ammonium acetate)	−	no	olive oil
[31]	lipids	5	tetrahydrofuran	−	MLR	kerogen extract
[32]	small bases	56	pH = 3/acetonitrile 50/50; 20/80; pH = 7/acetonitrile 50/50; 20/80	+	no	solvent
[33]	lipids	15	chloroform/methanol 1:2 1 % ammonia	−	no	solvent
[34]	peptide	4	0.1 % ammonium acetate/acetonitrile 0.5 % acetic acid 0.1 % ammonium acetate/acetonitrile	+	no	solvent
[35]	drug	99	ammonia (pH = 10)/methanol 50/50	+	PLS	solvent
[36]	glycans	8	1 mM NaOH/methanol 50/50	+	no	*Drosophila melanogaster*. solvent
[37]	steroid	7	10 mM ammonium acetate/acetonitrile/isopropanol (9/1); 0.02 % acetic acid/ acetonitrile/isopropanol 9/1	−	no	solvent; human plasma
[38]	natural products	4	water	+	MLR	secondary organic

Ref.	Compound type	Number of unique compounds	Eluent	Mode	Model	Matrix
						aerosol in water
[39]	small molecules	17	pH= 3.1 /acetonitrile	−/+	no	solvent
[40]	oligopeptide	12	1 mM ammonium acetate (pH = 6)/methanol 90/10; 1 mM ammonium acetate (pH = 10)/methanol 90/10	+	MLR, regression tree, SVR	solvent
[41]	glycans	13	0.1 % formic acid/acetonitrile 65/35	+	no	tryptic BSA peptides
[42]	lipophilic marine algal toxins	10	2 mM ammonium formate and 50 mM formic acid/acetonitrile (with additive) 40/60	+/−	no	solvent; mussel
[43]	flavonoids	6	formic acid acid (pH = 3)/acetonitrile	−	no	solvent
[44]	lipids	34	10 mM ammonium formate in water/acetonitrile (6/4) and 10 mM ammonium formate isopropanol/acetonitrile (9/1)	+/−	no	human plasma; mouse brain tissue; Jurkat cell pellet
[45]	small acids	25	0.025 % acetic acid/methanol	−	MLR	solvent
[46]	amino acids	17	2 % acetic acid/methanol 50/50	+	no	solvent
[47]	flavonoids	19	0.1 % formic acid/acetonitrile	−	no	extract of *G. biloba*

Table S2 log*IE* values collected from previous studies and measured in this study.

See logIE_values.xlsx

Table S3 The eluent compositions used for model development in ESI positive mode (* denotes measurements carried out during this study).

Organic modifier percentage	Organic modifier	Water phase percentage	Additive	Additive concentration (mM)	pH(aq)	Number of unique compounds	Ref.
100	Methanol	0	formic acid	27.0	2.7	35	*,[2]
100	Methanol	0	trifluoroacetic acid	5.0	3.6	11	[2]
100	Methanol	0	oxalic acid	5.0	4.1	11	[2]
100	Methanol	0	oxalic acid	1.0	5.4	11	[2]
100	Methanol	0	formic acid	5.0	5.4	11	[2]
100	Methanol	0	formic acid	1.0	5.9	11	[2]
100	Methanol	0	formic acid ammonium acetate	5.0	6.1	11	[2]
100	Methanol	0	formic acid ammonium acetate	5.0	6.3	11	[2]
100	Methanol	0	acetic acid	5.0	6.5	11	[2]
100	Methanol	0	formic acid	0.1	6.6	11	[2]
100	Methanol	0	formic acid ammonium acetate	5.0	7.6	11	[2]
100	Methanol	0	formic acid ammonium acetate	5.0	8.4	11	[2]
100	Methanol	0	ammonium formate	5.0	9.4	11	[2]
100	Methanol	0	ammonium acetate	5.0	10.1	11	[2]
100	Methanol	0	ammonia	5.0	12.2	11	[2]
90	Methanol	10	oxalic acid	1.0	2.1	20	[3]
90	Methanol	10	formic acid	3.0	2.7	20	[3]
90	Methanol	10	formic acid	1.0	2.9	20	[3]
90	Methanol	10	propionic acid	1.0	3.4	20	[3]
80	Methanol	20	formic acid	27.0	2.7	89	*
50	Methanol	50	trifluoroacetic acid	10.0	2.0	40	*
50	Methanol	50	formic acid	27.0	2.7	40	*
50	Methanol	50	ammonium fluoride	10.0	5.5	40	*

Organic modifier percentage	Organic modifier	Water phase percentage	Additive	Additive concentration (mM)	pH(aq)	Number of unique compounds	Ref.
50	Methanol	50	ammonium acetate	10.0	6.8	40	*
50	Methanol	50	ammonium formate	10.0	6.8	40	*
50	Methanol	50	ammonium bicarbonate	10.0	7.8	40	*
50	Methanol	50	ammonia	52.0	10.7	36	*
20	Methanol	80	formic acid	27.0	2.7	40	*
100	Acetonitrile	0	formic acid	27.0	2.7	35	*
90	Acetonitrile	10	oxalic acid	1.0	2.1	20	3
90	Acetonitrile	10	formic acid	1.0	2.9	20	3
90	Acetonitrile	10	propionic acid	1.0	3.4	20	3
80	Acetonitrile	20	trifluoroacetic acid	3.0	1.9	65	4
80	Acetonitrile	20	oxalic acid	1.0	2.1	17	5
80	Acetonitrile	20	formic acid	5.0	2.7	352	*, 3,6–9
80	Acetonitrile	20	ammonium acetate	1.0	3.0	20	4
80	Acetonitrile	20	acetic acid	3.0	3.2	17	5
80	Acetonitrile	20	ammonium acetate	1.0	3.5	20	4
80	Acetonitrile	20	ammonium acetate	1.0	4.0	20	4
80	Acetonitrile	20	ammonium acetate	1.0	4.5	19	4
80	Acetonitrile	20	ammonium acetate	1.0	5.0	26	4,7
80	Acetonitrile	20	ammonium acetate	1.0	5.5	20	4
80	Acetonitrile	20	–	–	–	16	5
80	Acetonitrile	20	ammonium acetate	1.0	6.0	20	4
80	Acetonitrile	20	ammonium acetate	1.0	6.5	20	4
80	Acetonitrile	20	ammonium acetate	1.0	6.7	15	4
80	Acetonitrile	20	ammonium acetate	1.0	7.0	46	4

Organic modifier percentage	Organic modifier	Water phase percentage	Additive	Additive concentration (mM)	pH(aq)	Number of unique compounds	Ref.
80	Acetonitrile	20	methylamine	0.0	7.9	14	[5]
80	Acetonitrile	20	ammonia	0.2	9.8	10	[7]
80	Acetonitrile	20	ammonia	10.0	10.7	36	[7]
50	Acetonitrile	50	trifluoroacetic acid	10.0	2.0	40	*
50	Acetonitrile	50	formic acid	27.0	2.7	45	*,[7]
50	Acetonitrile	50	ammonium acetate	2.5	5.0	10	[7]
50	Acetonitrile	50	ammonium fluoride	10.0	5.5	40	*
50	Acetonitrile	50	ammonium acetate	10.0	6.8	40	*
50	Acetonitrile	50	ammonium formate	10.0	6.8	40	*
50	Acetonitrile	50	ammonium bicarbonate	10.0	7.8	40	*
50	Acetonitrile	50	ammonia	0.5	9.8	10	[7]
50	Acetonitrile	50	ammonia	52.0	10.7	50	*
20	Acetonitrile	80	trifluoroacetic acid	10.0	1.9	28	[4]
20	Acetonitrile	80	formic acid	27.0	2.7	60	*,[7]
20	Acetonitrile	80	ammonium acetate	4.0	3.0	17	[4]
20	Acetonitrile	80	ammonium acetate	4.0	3.5	21	[4]
20	Acetonitrile	80	ammonium acetate	4.0	4.0	20	[4]
20	Acetonitrile	80	ammonium acetate	4.0	4.5	19	[4]
20	Acetonitrile	80	ammonium acetate	4.0	5.0	20	[4]
20	Acetonitrile	80	ammonium acetate	4.0	5.5	19	[4]
20	Acetonitrile	80	ammonium acetate	4.0	6.0	20	[4]
20	Acetonitrile	80	ammonium acetate	4.0	6.5	25	[4]
20	Acetonitrile	80	ammonium acetate	4.0	7.0	19	[4]

Organic modifier percentage	Organic modifier	Water phase percentage	Additive	Additive concentration (mM)	pH(aq)	Number of unique compounds	Ref.
90	Isopropanol	10	oxalic acid	1.0	2.1	20	[3]
90	Isopropanol	10	formic acid	1.0	2.9	18	[3]
90	Isopropanol	10	propionic acid	1.0	3.4	14	[3]
90	Acetone	10	oxalic acid	1.0	2.1	19	[3]
90	Acetone	10	formic acid	1.0	2.9	19	[3]
90	Acetone	10	propionic acid	1.0	3.4	20	[3]
0	–	100	oxalic acid	50.0	1.4	11	[2]
0	–	100	oxalic acid	5.0	2.1	11	[2]
0	–	100	trifluoroacetic acid	5.0	2.4	11	[2]
0	–	100	citric acid	5.0	2.5	11	[2]
0	–	100	formic acid	27.0	2.7	39	*,[2]
0	–	100	oxalic acid	1.0	2.8	11	[2]
0	–	100	formic acid	5.0	2.8	11	[2]
0	–	100	formic acid ammonium acetate	5.0	3.0	11	[2]
0	–	100	formic acid. ammonium formate	5.0	3.0	11	[2]
0	–	100	formic acid	1.0	3.1	11	[2]
0	–	100	acetic acid	5.0	3.2	11	[2]
0	–	100	formic acid ammonium acetate	5.0	3.3	11	[2]
0	–	100	formic acid ammonium acetate	5.0	3.5	11	[2]
0	–	100	formic acid	0.1	3.7	11	[2]
0	–	100	formic acid ammonium acetate	5.0	4.0	11	[2]
0	–	100	formic acid. ammonium formate	5.0	4.0	11	[2]
0	–	100	formic acid ammonium acetate	5.0	5.0	11	[2]

Organic modifier percentage	Organic modifier	Water phase percentage	Additive	Additive concentration (mM)	pH(aq)	Number of unique compounds	Ref.
0	–	100	formic acid ammonium acetate	5.0	5.0	11	[2]
0	–	100	ammonium formate	5.0	5.5	11	[2]
0	–	100	formic acid ammonium acetate	5.0	5.9	11	[2]
0	–	100	formic acid ammonium formate	5.0	5.9	11	[2]
0	–	100	ammonium acetate	1.0	6.1	11	[2]
0	–	100	ammonium acetate	5.0	6.7	11	[2]
0	–	100	ammonium acetate	10.0	6.8	11	[2]
0	–	100	methylamine	1.0	7.9	9	[2]
0	–	100	ammonia ammonium acetate	5.0	8.1	11	[2]
0	–	100	ammonia ammonium acetate	5.0	9.0	11	[2]
0	–	100	ammonia	1.0	9.4	11	[2]
0	–	100	ammonia	5.0	10.1	11	[2]
0	–	100	ammonia	10.0	10.3	11	[2]

Table S4 The eluent compositions used for model development in ESI negative mode.

Organic modifier percentage	Organic modifier	water phase percentage	Additive	additive concentration (mM)	pH(aq)	Number of unique compounds	Ref.
100	Methanol	0	ammonia	52.0	10.5	17	*
100	Methanol	0	-	-	7.0	17	*
80	Methanol	20	ammonia	10.0	10.5	17	*
80	Methanol	20	-	-	7.0	16	*
60	Methanol	40	ammonia	21.0	10.5	17	*
60	Methanol	40	-	-	7.0	17	*
40	Methanol	60	ammonia	31.0	10.5	17	*
40	Methanol	60	-	-	7.0	17	*
20	Methanol	80	ammonia	41.0	10.5	17	*
20	Methanol	80	-	-	7.0	17	*
0	Methanol	100	-	-	7.0	17	*
100	Acetonitrile	0	ammonia	52.0	10.5	60	[10]
100	Acetonitrile	0	-	-	7.0	17	[11]
80	Acetonitrile	20	ammonia	10.0	10.5	101	*, [8,11–13]
80	Acetonitrile	20	ammonium acetate	1.0	7.8	59	[14]
80	Acetonitrile	20	ammonium acetate	1.0	7.0	22	*
80	Acetonitrile	20	-	-	7.0	16	[11]
80	Acetonitrile	20	ammonium acetate	0.2	5.0	60	[14]
80	Acetonitrile	20	ammonium acetate	1.0	5.0	62	[14]
80	Acetonitrile	20	ammonium acetate	1.0	3.5	56	[14]
80	Acetonitrile	20	formic acid	5.0	2.8	61	[14]
60	Acetonitrile	40	ammonia	21.0	10.5	17	[11]
60	Acetonitrile	40	-	-	7.0	17	[11]
50	Acetonitrile	50	ammonia	26.0	10.5	59	[14]
40	Acetonitrile	60	ammonia	31.0	10.5	17	[11]
40	Acetonitrile	60	-	-	7.0	17	[11]
20	Acetonitrile	80	ammonia	41.0	10.5	62	[14]
20	Acetonitrile	80	-	-	7.0	17	[11]
20	Acetonitrile	80	ammonium acetate	4.0	5.0	57	[14]
20	Acetonitrile	80	formic acid	21.0	2.8	49	[14]
0	-	100	ammonia	52.0	10.5	17	[11]

Organic modifier percentage	Organic modifier	water phase percentage	Additive	additive concentration (mM)	pH(aq)	Number of unique compounds	Ref.
0	-	100	ammonia	52.0	10.5	17	[11]
0	-	100	-	-	7.0	17	[11]

Table S5 Classifications of studied compounds using ClassyFire.[48]

See classyfire_classification.csv

Table S6 The most prominent superclasses covered by the compounds included in this study. Classifications of studied compounds using ClassyFire.[48]

Superclass	n
Benzenoids	163
Organoheterocyclic compounds	83
Organic acids and derivatives	54
Organic nitrogen compounds	28
Lipids and lipid-like molecules	16
Organic oxygen compounds	9
Phenylpropanoids and polyketides	6
Organosulfur compounds	5
Alkaloids and derivatives	4
Organophosphorus compounds	3

Table S7 Compounds used in validation and application study. P- denotes pesticide, M- denotes mycotoxin and T -transformation compound.

#	CAS	Name	SMILES
M1	88337-96-6	deoxynivalenol-15-acetate	CC1=C[C@@H]2[C@]([C@@H](C1=O)O)([C@]3(C[C@H]([C@H]([C@@]34CO4)O2)O)C)COC(=O)C
P2	30614-22-3	pirimicarb-desmethyl	CC1=C(N=C(N=C1OC(=O)N(C)C)NC)C
P3	100784-20-1	halosulfuron-methyl	CN1C(=C(C(=N1)Cl)C(=O)OC)S(=O)(=O)NC(=O)NC2=NC(=CC(=N2)OC)OC
P4	2310-17-0	phosalone	CCOP(=S)(OCC)SCN1C2=C(C=C(C=C2)Cl)OC1=O
P5	950-35-6	paraoxon-methyl	COP(=O)(OC)OC1=CC=C(C=C1)[N+](=O)[O-]
P6	107-49-3	TEPP	CCOP(=O)(OCC)OP(=O)(OCC)OCC
P7	31972-43-7	fenamiphos-sulfoxide	CCOP(=O)(NC(C)C)OC1=CC(=C(C=C1)S(=O)C)C
P8	3761-41-9	fenthion-sulfoxide	CC1=C(C=CC(=C1)OP(=S)(OC)OC)S(=O)C
P9	72490-01-8	fenoxycarb	CCOC(=O)NCCOC1=CC=C(C=C1)OC2=CC=CC=C2
P10	57018-04-9	tolclofos-methyl	CC1=CC(=C(C(=C1)Cl)OP(=S)(OC)OC)Cl
P11	88671-89-0	myclobutanil	CCCCC(CN1C=NC=N1)(C#N)C2=CC=C(C=C2)Cl
P12	67129-08-2	metazachlor	CC1=C(C(=CC=C1)C)N(CN2C=CC=N2)C(=O)CCl
P13	61213-25-0	flurochloridone	C1C(C(C(=O)N1C2=CC=CC(=C2)C(F)(F)F)Cl)CCl
P14	42874-03-3	oxyfluorfen	CCOC1=C(C=CC(=C1)OC2=C(C=C(C=C2)C(F)(F)F)Cl)[N+](=O)[O-]
P15	34256-82-1	acetochlor	CCC1=CC=CC(=C1N(COCC)C(=O)CCl)C
P16	63284-71-9	nuarimol	C1=CC=C(C(=C1)C(C2=CC=C(C=C2)F)(C3=CN=CN=C3)O)Cl
P17	79983-71-4	hexaconazole	CCCCC(CN1C=NC=N1)(C2=C(C=C(C=C2)Cl)Cl)O
M18	23452-05-3	alternariol-monomethyl ether	CC1=CC(=CC2=C1C3=CC(=CC(=C3C(=O)O2)O)OC)O
P19	55512-33-9	pyridate	CCCCCCCCSC(=O)OC1=CC(=NN=C1C2=CC=CC=C2)Cl
P20	86-86-2	1-naphthylacetamide	C1=CC=C2C(=C1)C=CC=C2CC(=O)N
P21	175013-18-0	pyraclostrobin	COC(=O)N(C1=CC=CC=C1COC2=NN(C=C2)C3=CC=C(C=C3)Cl)OC
M22	1165-39-5	aflatoxin G1	COC1=C2C3=C(C(=O)OCC3)C(=O)OC2=C4C5C=COC5OC4=C1
P23	74115-24-5	clofentezine	C1=CC=C(C(=C1)C2=NN=C(N=N2)C3=CC=CC=C3Cl)Cl
P24	112281-77-3	tetraconazole	C1=CC(=C(C=C1Cl)Cl)C(CN2C=NC=N2)COC(C(F)F)(F)F
P25	81777-89-1	clomazone	CC1(CON(C1=O)CC2=CC=CC=C2Cl)C
P26	24579-73-5	propamocarb	CCCOC(=O)NCCCN(C)C

#	CAS	Name	SMILES
P27	119168-77-3	tebufenpyrad	CCC1=NN(C(=C1Cl)C(=O)NCC2=CC=C(C=C2)C(C)(C)C)C
P28	135410-20-7	acetamiprid	CC(=NC#N)N(C)CC1=CN=C(C=C1)Cl
P29	10265-92-6	methamidophos	COP(=O)(N)SC
P30	173584-44-6	indoxacarb	COC(=O)[C@]12CC3=C(C1=NN(CO2)C(=O)N(C4=CC=C(C=C4)OC(F)(F)F)C(=O)OC)C=CC(=C3)Cl
P31	77732-09-3	oxadixyl	CC1=C(C(=CC=C1)C)N(C(=O)COC)N2CCOC2=O
P32	119-12-0	pyridaphenthion	CCOP(=S)(OCC)OC1=NN(C(=O)C=C1)C2=CC=CC=C2
P33	470-90-6	chlorfenvinphos	CCOP(=O)(OCC)O/C(=C\Cl)/C1=C(C=C(C=C1)Cl)Cl
M34	2270-40-8	diacetoxyscirpenol	CC1=C[C@@H]2[C@](CC1)([C@]3([C@@H]([C@H]([C@H]([C@]34CO4)O2)O)OC(=O)C)C)COC(=O)C
M35	1162-65-8	aflatoxin B1	COC1=C2C3=C(C(=O)CC3)C(=O)OC2=C4[C@@H]5C=CO[C@@H]5OC4=C1
T1	620-08-6	4-methoxypyridine	COc1ccncc1
T2	271-44-3	indazole	[nH]1ncc2ccccc12
T3	123-08-0	4-hydroxybenzaldehyde	Oc1ccc(C=O)cc1
T4	66-40-0	tetraethylammonium	CC[N+](CC)(CC)CC
T5	123-11-5	*p*-anisaldehyde	COc1ccc(C=O)cc1
T6	934-00-9	3-methoxycatechol	COc1cccc(O)c1O
T7	102-69-2	tripropylamine	CCCN(CCC)CCC
T8	56-85-9	glutamine	N[C@@H](CCC(N)=O)C(O)=O
T9	140-10-3	cinnamic acid	OC(=O)C=Cc1ccccc1
T10	15799-79-8	3-methoxy-*N,N*-dimethylaniline	COc1cccc(c1)N(C)C
T11	100-22-1	4-dimethylamino-*N,N*-dimethylaniline	CN(C)c1ccc(cc1)N(C)C
T12	88-99-3	phthalic acid	OC(=O)c1ccccc1C(O)=O
T13	66-71-7	1,10-phenanthroline	c1cnc2c(c1)ccc3cccnc23
T14	13010-31-6	tetrapropylammonium	CCC[N+](CCC)(CCC)CCC
T15	140-40-9	2-acetamido-5-nitrothiazole	CC(=O)Nc1sc(cn1)[N+]([O-])=O
T16	41394-05-2	metamitrone	CC1=NN=C(c2ccccc2)C(=O)N1N

#	CAS	Name	SMILES
T17	34123-59-6	isoproturon	CC(C)c1ccc(NC(=O)N(C)C)cc1
T18	1646-88-4	aldicarb-sulfone	CNC(=O)O\N=C\C(C)(C)[S](C)(=O)=O
T19	15972-60-8	alachlor	CCc1cccc(CC)c1N(COC)C(=O)CCl
T20	13067-93-1	cyanophenphos	CCO[P](=S)(Oc1ccc(cc1)C#N)c2ccccc2
T21	51-34-3	scopolamine	CN1[C@@H]2CC(C[C@H]1[C@H]3[C@@H]2O3)OC(=O)[C@H](CO)C4=CC=CC=C4
T22	2597-03-7	phenthoate	CCOC(=O)C(S[P](=S)(OC)OC)c1ccccc1
T23	23564-05-8	thiophanate-methyl	COC(=O)NC(=S)NC1=CC=CC=C1NC(=S)NC(=O)OC
T24	112281-77-3	tetraconazole	FC(F)C(F)(F)OCC(Cn1cncn1)c2ccc(Cl)cc2Cl
T25	417706-59-3	4-CF_3-C_6H_4-P(pyrr)	FC(F)(F)c1ccc(N=P(N2CCCC2)(N2CCCC2)N2CCCC2)cc1
T26	2097489-41-1	4-[2-(4-nitrophenyl)diazenyl]-*N*-(phenyldi-1-pyrrolidinylphosphoranylidene)benzenamine	O=[N+]([O-])c1ccc(N=Nc2ccc(N=P(c3ccccc3)(N3CCCC3)N3CCCC3)cc2)cc1
T27	2097489-43-3	4-[2-[4-[(diphenyl-1-pyrrolidinylphosphoranylidene)amino]phenyl]diazenyl]-*N,N*-dimethyl-benzenamine	CN(C)c1ccc(N=Nc2ccc(N=P(c3ccccc3)(c3ccccc3)N3CCCC3)cc2)cc1
T28	2097489-43-3	*N,N*-dimethyl-4-[2-[4-[(triphenylphosphoranylidene)amino]phenyl]diazenyl]-benzenamine	CN(C)c1ccc(N=Nc2ccc(N=P(c3ccccc3)(c3ccccc3)N3CCCC3)cc2)cc1
T29	33354-65-3	2-Cl-C_6H_4-P_2(pyrr)	CN(C)c1ccc(N=Nc2ccc(N=P(c3ccccc3)(c3ccccc3)c3ccccc3)cc2)cc1
T30	417706-58-2	Phe-Phe-Phe-Phe	Clc1ccccc1N=P(N=P(N1CCCC1)(N1CCCC1)N1CCCC1)(N1CCCC1)N1CCCC1
T31	2667-02-9	reserpine	NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O

Table S8 Instruments used to study ionization efficiencies in this study.

Usage	Source	Mass analyzer	Name	Vendor	Lab
Ionization efficiency model development	ESI	Ion trap	Agilent XCT	Agilent	University of Tartu
Ionization efficiency model development	ESI	Linear ion trap	Thermo LTQ	Thermo Fisher Scientific	Janssen Pharmaceutica
Ionization efficiency model development	HESI−II	Linear ion trap	Thermo LTQ	Thermo Fisher Scientific	Janssen Pharmaceutica
Ionization efficiency model development	Z-spray	Quadrupole time-of-flight	Waters Synapt G.2	Waters	Janssen Pharmaceutica
Ionization efficiency model development	JetStream	Single quadrupole	Agilent Single Quad 6100	Agilent	University Lyon 1
Ionization efficiency model development	ESI	Triple quadrupole	Agilent 6495	Agilent	University of Tartu
Ionization efficiency model development/ Validation	JetStream	Triple quadrupole	Agilent 6495	Agilent	University of Tartu
Ionization efficiency model development	TurboSpray	Triple quadrupole	Sciex API 4000	Sciex	Janssen Pharmaceutica
Ionization efficiency model development	ESI	Triple quadrupole	Varian J-320	Varian	University of Tartu

Table S9 Significant descriptors in ESI positive mode model.

Descriptor type	*N*	Descriptor names
Acidic group count	1	nAcid
AlogP[49]	2	ALogP, ALogp2
Atom count	5	nC, nH, nHeavyAtom, nN, nX
Autocorrelation[50]	208	AATS0i, AATS0p, AATS1e, AATS1i, AATS1p, AATS1s, AATS1v, AATS2i, AATS2p, AATS2s, AATS3e, AATS3i, AATS3m, AATS3s, AATS4e, AATS4i, AATS4m, AATS4p, AATS4s, AATS5i, AATS5s, AATS5v, AATS6i, AATS6p, AATS6s, AATS7i, AATS7m, AATS8e, AATS8i, AATS8m, AATS8p, AATS8v, AATSC0e, AATSC0i, AATSC0p, AATSC1e, AATSC1i, AATSC1m, AATSC1p, AATSC1s, AATSC1v, AATSC2e, AATSC2i, AATSC2m, AATSC2s, AATSC2v, AATSC3c, AATSC3e, AATSC3i, AATSC3s, AATSC3v, AATSC4c, AATSC4e, AATSC4i, AATSC4m, AATSC4s, AATSC4v, AATSC5i, AATSC5m, ATSC5v, AATSC6e, AATSC6i, AATSC6s, AATSC6v, AATSC7e, AATSC7i, AATSC7m, AATSC7v, AATSC8e, AATSC8i, AATSC8m, AATSC8p, AATSC8s, AATSC8v, ATS0e, ATS0i, ATS0m, ATS0s, ATS0v, ATS1i, ATS1s, ATS2e, ATS2s, ATS2v, ATS3i, ATS3s, ATS4i, ATS4p, ATS5m, ATS8e, ATS8i, ATS8s, ATSC0i, ATSC0m, ATSC0p, ATSC1c, ATSC1e, ATSC1m, ATSC1p, ATSC1s, ATSC2c, ATSC2e, ATSC2i, ATSC2m, ATSC2s, ATSC3c, ATSC3e, ATSC3m, ATSC3p, ATSC3v, ATSC4c, ATSC4e, ATSC4i, ATSC4m, ATSC4p, ATSC4s, ATSC4v, ATSC5c, ATSC5i, ATSC5m, ATSC5p, ATSC5s, ATSC6c, ATSC6e, ATSC6m, ATSC6p, ATSC6s, ATSC6v, ATSC7i, ATSC7m, ATSC8c, ATSC8e, ATSC8i, ATSC8s, ATSC8v, GATS1c, GATS1e, GATS1i, GATS1m, GATS1p, GATS1s, GATS2c, GATS2e, GATS2i, GATS2m, GATS2p, GATS2s, GATS2v, GATS3c, GATS3m, GATS3p, GATS3s, GATS3v, GATS4c, GATS4e, GATS4m, GATS4p, GATS4s, GATS4v, GATS5c, GATS5e, GATS5i, GATS5m, GATS5v, GATS6e, GATS6p, GATS6s, GATS6v, GATS7m, GATS8c, GATS8e, GATS8i, GATS8p, GATS8v, MATS1i, MATS1p, MATS1s, MATS2c, MATS2e, MATS2m, MATS2v, MATS3c, MATS3e, MATS3i, MATS3m, MATS3s, MATS3v, MATS4e, MATS4i, MATS4m, MATS4p, MATS4s, MATS4v, MATS5c, MATS5e, MATS5m, MATS5p, MATS5s, MATS5v, MATS6e, MATS6p, MATS6s, MATS7c, MATS8c, MATS8i, MATS8m, MATS8s, MATS8v
Barysz matrix[50]	41	EE_Dzi, EE_Dzm, EE_Dzp, EE_Dzs, EE_DzZ, SM1_Dzi, SM1_Dzp, SM1_Dzs, SM1_DzZ, SM1_Dzv, SpAD_Dzp, SpAD_DzZ, SpMax_Dze, SpMax_Dzi, VE1_Dzi, VE1_Dzm, VE1_Dzp, VE1_Dzs, VE1_DzZ, VE1_Dzv, VE2_Dze, VE2_Dzi, VE2_Dzs, VE2_DzZ, VE2_Dzv, VE3_Dzp, VE3_Dzs, VE3_DzZ, VE3_Dzv, VR1_Dzi, VR1_Dzm, VR1_Dzp, VR1_Dzs, VR1_DzZ, VR1_Dzv, VR2_Dzi, VR2_Dzp, VR2_Dzs, VR2_DzZ, VR2_Dzv, VR3_Dzs
Burden modified eigenvalues[50]	55	SpMax1_Bhs, SpMax2_Bhe, SpMax2_Bhv, SpMax3_Bhe, SpMax3_Bhi, SpMax3_Bhm, SpMax3_Bhp, SpMax4_Bhe, SpMax4_Bhi, SpMax4_Bhs, SpMax5_Bhp, SpMax5_Bhv, SpMax6_Bhe, SpMax6_Bhi,

Descriptor type	*N*	Descriptor names
		SpMax6_Bhp, SpMax6_Bhv, SpMax7_Bhm, SpMax7_Bhp, SpMax7_Bhs, SpMax8_Bhe, SpMax8_Bhm, SpMax8_Bhp, SpMax8_Bhs, SpMin1_Bhi, SpMin1_Bhm, SpMin1_Bhp, SpMin1_Bhs, SpMin1_Bhv, SpMin2_Bhi, SpMin2_Bhp, SpMin2_Bhs, SpMin2_Bhv, SpMin3_Bhe, SpMin3_Bhi, SpMin3_Bhm, SpMin3_Bhs, SpMin3_Bhv, SpMin4_Bhe, SpMin4_Bhi, SpMin4_Bhs, SpMin4_Bhv, SpMin5_Bhe, SpMin5_Bhs, SpMin5_Bhv, SpMin6_Bhe, SpMin6_Bhi, SpMin6_Bhv, SpMin7_Bhm, SpMin7_Bhp, SpMin7_Bhs, SpMin7_Bhv, SpMin8_Bhe, SpMin8_Bhi, SpMin8_Bhp, SpMin8_Bhv
Constitutional	4	Mi, Mse, Mv, Sv
Crippen[51]	1	CrippenLogP
Detour matrix[50]	6	EE_Dt, VE1_Dt, VE2_Dt, VE3_Dt, VR1_Dt, VR2_Dt
Atom type electrotopological state[52]	67	DELS, DELS2, gmax, gmin, hmin, maxaaaC, MAXDN, maxdNH, MAXDP, MAXDP2, maxdsCH, maxdssC, maxHAvin, maxHBint4, maxHBint5, maxHCHnX, maxHdCH2, maxHdsCH, maxHsNH2, maxHsSH, maxsNH3p, maxsOH, maxssPH, maxssssBm, maxtN, maxwHBa, maxwHBd, minaaN, minaaS, mindO, mindsCH, mindsN, mindssC, minHBa, minHBd, minHBint5, minHdsCH, minsAsH2, minsNH2, minsOH, minsSeH, minssNH, minsssCH, minssSe, minsssN, minwHBa, nddssSe, ndO, nHsOH, nsCH3, nsSnH3, nssO, nssSe, SaaCH, SdO, SdS, SdssC, SdsssP, SHBa, SHBint9, SHdsCH, SHsNH2, SsNH2, SsNH3p, SsSnH3, SsssCH, sumI
Eluent	5	NH4, pH_aq, polarity_index, surface_tension, viscosity
Extended topochemical atom[53]	18	ETA_Alpha, ETA_Beta, ETA_Beta_ns, ETA_Beta_s, ETA_BetaP, ETA_dAlpha_B, ETA_dBeta, ETA_dBetaP, ETA_dEpsilon_A, ETA_Epsilon_5, ETA_Eta_B, ETA_Eta_B_RC, ETA_Eta_F_L, ETA_Eta_L, ETA_EtaP_B, ETA_EtaP_B_RC, ETA_EtaP_F_L, ETA_Shape_Y
Hybridization ratio	1	HybRatio
Information content[50]	24	BIC0, BIC1, BIC3, BIC4, BIC5, CIC0, CIC1, CIC3, CIC4, CIC5, IC1, IC2, IC3, IC4, MIC0, MIC4, SIC3, SIC4, SIC5, ZMIC1, ZMIC2, ZMIC4, TIC1, TIC5
Molecular distance edge[54]	7	MDEC-11, MDEC-14, MDEC-24, MDEC-34, MDEN-13, MDEN-22, MDEN-33
Molecular linear free energy relation[55]	3	MLFER_BH, MLFER_E, MLFER_S
Bond count	1	nBondsM
Carbon types	1	C2SP1

Table S10 Significant descriptors in ESI negative mode model.

Descriptor type	*N*	Descriptor names
Atom count	1	nX
Autocorrelation[50]	57	AATS1e, AATS1i, AATS2p, AATS2v, AATS3e, AATS4e, AATS5p, AATS8s, AATSC0c, AATSC1p, AATSC3c, AATSC3p, AATSC3v, AATSC4c, AATSC5p, AATSC6v, AATSC7p, ATS0m, ATS0s, ATS1m, ATS1s, ATS3i, ATS3m, ATS3s, ATS4m, ATSC0s, ATSC0v, ATSC1v, ATSC3c, ATSC3p, ATSC4e, ATSC8v, GATS1e, GATS1v, GATS2e, GATS3c, GATS3i, GATS4m, GATS4v, GATS5i, GATS5m, GATS5s, GATS6v, GATS8c, GATS8m, GATS8p, MATS1m, MATS1p, MATS2e, MATS2s, MATS2v, MATS3i, MATS3m, MATS3v, MATS6m, MATS8i, MATS8s
Barysz matrix[50]	32	EE_Dzm, EE_DzZ, SM1_Dze, SM1_Dzi, SM1_Dzm, SM1_Dzs, SM1_Dzv, SpAD_Dzp, SpAD_Dzv, SpDiam_Dze, SpMAD_DzZ, VE1_Dzi, VE1_Dzp, VE1_DzZ, VE2_Dze, VE2_Dzi, VE2_Dzp, VE2_DzZ, VE2_Dzv, VE3_Dze, VE3_Dzi, VE3_Dzm, VE3_Dzp, VE3_Dzs, VE3_DzZ, VR1_Dze, VR1_Dzi, VR1_Dzp, VR1_DzZ, VR1_Dzv, VR3_Dzm, VR3_Dzs
BCUT[56]	1	BCUTc-1l
Burden modified eigenvalue[50]	19	SpMax1_Bhm, SpMax1_Bhp, SpMax2_Bhm, SpMax2_Bhp, SpMax2_Bhs, SpMax3_Bhs, SpMax7_Bhe, SpMax7_Bhi, SpMax8_Bhp, SpMin1_Bhe, SpMin1_Bhi, SpMin3_Bhp, SpMin3_Bhs, SpMin4_Bhs, SpMin4_Bhv, SpMin5_Bhe, SpMin5_Bhi, SpMin6_Bhm, SpMin7_Bhm
Chi Chain[57]	1	VCH-6
Detour matrix[50]	1	VE3_Dt
Atom type electrotopological state[52]	7	DELS2, maxdsCH, minHCsats, nHBint4, SddssS, SdsCH, SHBd
Eluent	4	pH_aq, polarity_index, surface_tension, viscosity
Extended topochemical atom[53]	6	ETA_BetaP, ETA_dBeta, ETA_Eta_L, ETA_Eta_R, ETA_EtaP_F_L, ETA_Shape_P
Information content[50]	8	BIC2, BIC3, BIC4, BIC5, CIC4, MIC3, SIC1, ZMIC3
Molecular linear free energy relation[55]	1	MLFER_A
Chi path[57]	1	AVP-0
Weighted path[58]	1	WTPT-2
Topological charge[1]	2	GGI3, GGI4
Topological distance matrix[50]	2	VE1_D, VE2_D
Wiener numbers[59]	1	WPATH

Table S11 Ionization efficiency prediction model characteristics: absolute error distribution in log*IE* units and times differences.

	positive model				negative model			
	training set		test set		training set		test set	
	log*IE* units	times difference	log*IE* units	times difference	log*IE* units	times difference	log*IE* units	times difference
RMSE	0.29	1.94	0.48	3.01	0.29	1.97	0.35	2.25
min	1.83E-05	1.00	5.61E-04	1.00	1.93E-04	1.00	7.71E-03	1.02
Q1	0.07	1.17	0.12	1.31	0.07	1.19	0.10	1.27
median (Q2)	0.16	1.43	0.24	1.76	0.16	1.43	0.24	1.74
average	0.21	1.61	0.34	2.18	0.21	1.61	0.28	1.90
Q3	0.29	1.93	0.45	2.81	0.27	1.87	0.39	2.43
max	1.97	92.94	2.51	323.67	2.15	139.89	1.05	11.10

Table S12 TOP 15 most important features in developed models in ESI positive and negative mode.

#	Descriptor names	
	ESI+	**ESI−**
1	CIC0	SM1_Dzm
2	ATS4i	GGI4
3	nH	DELS2
4	ATS3i	pH_aq
5	pH_aq	ETA_Eta_R
6	ATS4p	SM1_Dze
7	hmin	ATS1m
8	nN	ATS4m
9	viscosity	ATS0m
10	AATS2s	viscosity
11	ATS1i	surface_tension
12	SpMin4_Bhs	ATS0s
13	NH4	MIC3
14	SpMin5_Bhs	polarity_index
15	surface_tension	ATSC0s

Table S13 Absolute concentration prediction error characteristics based on the validation set.

	times difference
min	1.00
Q1 (25%)	1.38
median (Q2)	2.06
average	5.38
Q3 (50%)	4.41
max	93.55

Table S14 Comparison of errors in concentration prediction by compound in case of pesticides and mycotoxins in cereal matrices using three different approaches.

compound	**retention time**	**mean error** (using predicted ionization efficiency)	**mean error** (using 6 compounds[1] as internal standards and for quantifying using the response factor of the retention time wise nearest internal standard)	**mean error** (assuming equal response factors and using acetamiprid as internal standard)
methamidophos	0.85	1.75	12.60	7.34
propamocarb	1.23	7.34	2.80	4.80
pirimicarb-desmethyl	1.71	14.71	1142.42	5.92
deoxynivalenol-15-acetate	2.48	51.82	29.69	62.79
acetamiprid	2.79	4.08	2.15	1.47
TEPP	3.10	1.28	1.56	1.52
1-naphthylacetamide	3.14	1.67	1.51	2.55
fenamiphos-sulfoxide	3.15	1.26	2.23	1.26
diacetoxyscirpenol	3.37	10.29	5.01	10.37
aflatoxin G1	3.37	1.46	1.48	2.16
oxadixyl	3.39	1.36	1.35	2.26
paraoxon-methyl	3.43	4.29	1.50	2.94
aflatoxin B1	3.58	1.83	232.94	1.65
fenthion-sulfoxide	3.84	2.57	168.63	1.37
metazachlor	4.34	1.57	320.97	1.86
clomazone	4.45	1.81	169.02	1.50
nuarimol	4.51	1.30	309.76	1.82
alternariol-mono	4.74	14.55	7916.96	45.97
myclobutanil	4.88	1.42	160.12	1.20
halosulfuron-methy	4.93	3.06	283.66	1.65
tetraconazole	5.02	1.78	177.02	1.17
pyridaphenthion	5.03	2.15	135.82	1.34
hexaconazole	5.21	1.47	246.25	1.59
acetochlor	5.28	2.50	897.34	5.21
fenoxycarb	5.28	2.16	437.50	2.77
flurochloridone	5.30	1.79	3718.81	21.59
chlorfenrinphos	5.45	6.32	30.49	1.90
pyraclostrobin	5.84	3.60	5.32	4.09
clofenfezine	5.95	1.40	74.42	4.17
tolclofos-methyl	5.96	2.65	323.54	18.13

compound	retention time	mean error (using predicted ionization efficiency)	mean error (using 6 compounds[1] as internal standards and for quantifying using the response factor of the retention time wise nearest internal standard)	mean error (assuming equal response factors and using acetamiprid as internal standard)
phosalone	5.98	3.33	212.02	11.88
indoxacarb	6.05	8.38	14.29	1.32
tebufenpyrad	6.22	1.70	18.62	1.17
oxyfluorfen	6.42	18.86	1285.53	72.02
pyridate	7.49	3.58	49.44	3.03
Mean error		5.46	525.51	8.96

[1]6 internal standards and corresponding retention times:

- glutamine 0.46
- scopolamine 1.69
- 3-methoxycatechol 2.02
- Phe-Phe-Phe-Phe 3.58
- *N,N*-dimethyl-4-[2-[4-[(triphenylphosphoranylidene)amino]phenyl]diazenyl]-benzenamine 5.33
- 2-Cl-C_6H_4-P_2(pyrr) 7.01

Table S15 Comparison of errors in concentration prediction by matrix in case of pesticides and mycotoxins in cereal matrices using ionization efficiency prediction.

matrix	**average error**
oat	6.3
barley	5.7
maize	5.4
rye	5.4
wheat	5.4
rice	4.8
solvent	4.6

Table S16 Properties of pesticides studied in application study.

See chemical_properties_35_pesticides.xlsx

Table S17 Raw data collected in application study in example of pesticides and mycotoxins in cereals.

See cereals_raw_data.csv

Table S18 Summary of data used for model development and concentration prediction.

	Mode	**ESI+**	**ESI−**
Ionization efficiency model development	# of unique compounds	353	101
	# of eluent compositions	106	33
	# of instruments	7	3
	# of labs	3	2
	Methods used	direct infusion, flow injection analysis	direct infusion, flow injection analysis
	Total of log*IE* values	**3139**	**1286**
Concentration prediction (validation)	# of analytes	35	
	type of analytes	pesticides, mycotoxins	
	the lowest concentration	$3.6 \cdot 10^{-9}$ M	
	the highest concentration	$3.5 \cdot 10^{-4}$ M	
	measurement type	LC/ESI/MS gradient elution	
	lab	Tartu	
	sample	cereal	
	# of samples	6	
	Total of datapoints	**2233**	

Figure S1 Comparison of the chemical space coverage based on logP values. a – ESI positive mode; b – ESI negative mode. The comparison is made with databases of DrugBank,[60] Human Metabolome Database (HMDB),[61] data used in model development, and data pooled from literature (Table S1) HMDB also includes compounds that have not been and cannot be measured with LC/MS.

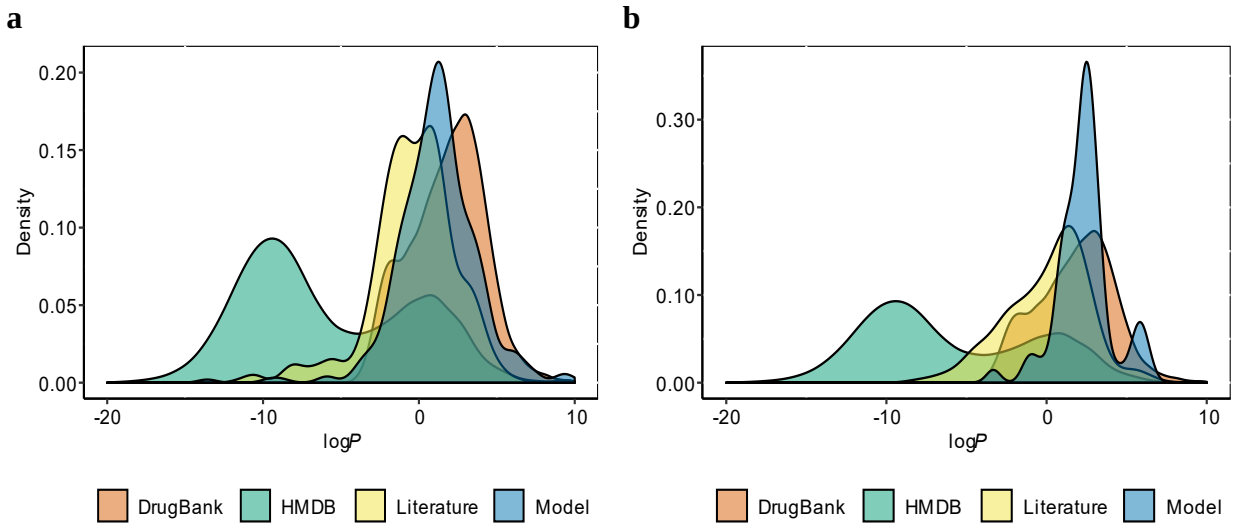


Figure S2 A comparison of the chemical space covered by compounds included in this study (training, test, and validation sets) in comparison to (left) NORMAN, (middle) HMDB, and (right) DrugBank databased. Each dot represents one compound. The compounds from current study are marked with blue and compounds from suspect screening databases are marked with yellow. For NORMAN and HMDB, only the compounds that have LC/MS records available were used. It should be noted that Drugbank also includes proteins and other larger compounds not covered by the proposed method.

The principal component analysis was conducted based on the PaDEL descriptor, and shown are the scores plots from the first two principal components. In all cases first and second principal component explain relatively small part of the total variance (21 to 29%). This is expected, as the chemical space has high dimensionality and includes compounds from very different classes.

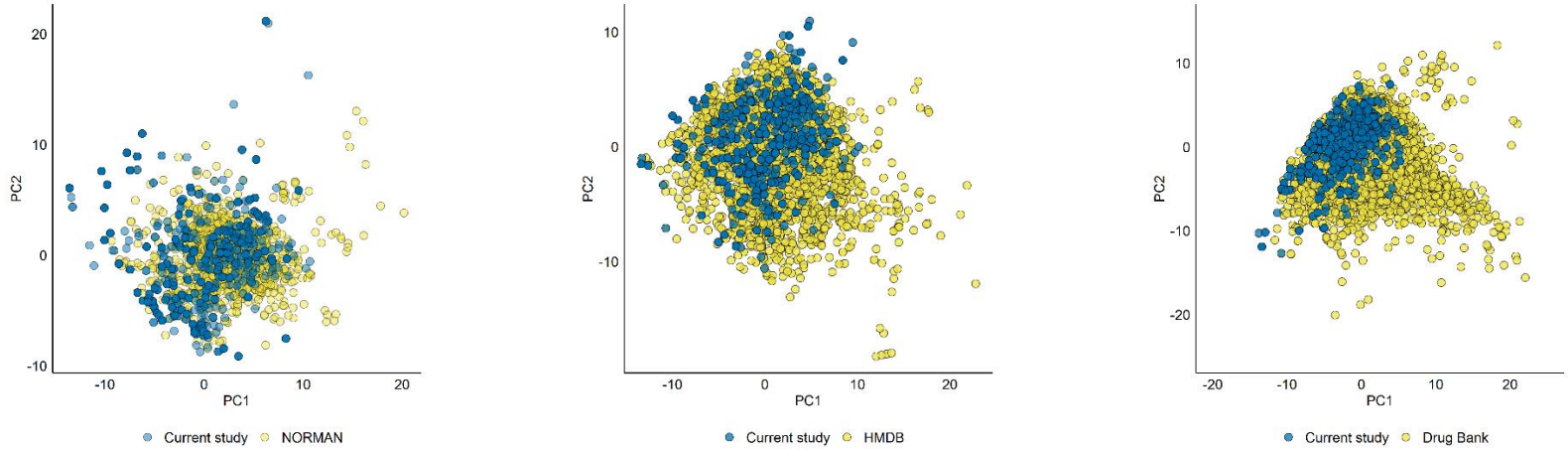


Figure S3 The PCA analysis of the compounds from training, test and validation set based on the PaDEL descriptors. First two principal component explain ca 30% of the total variance. Each dot represents one compound.

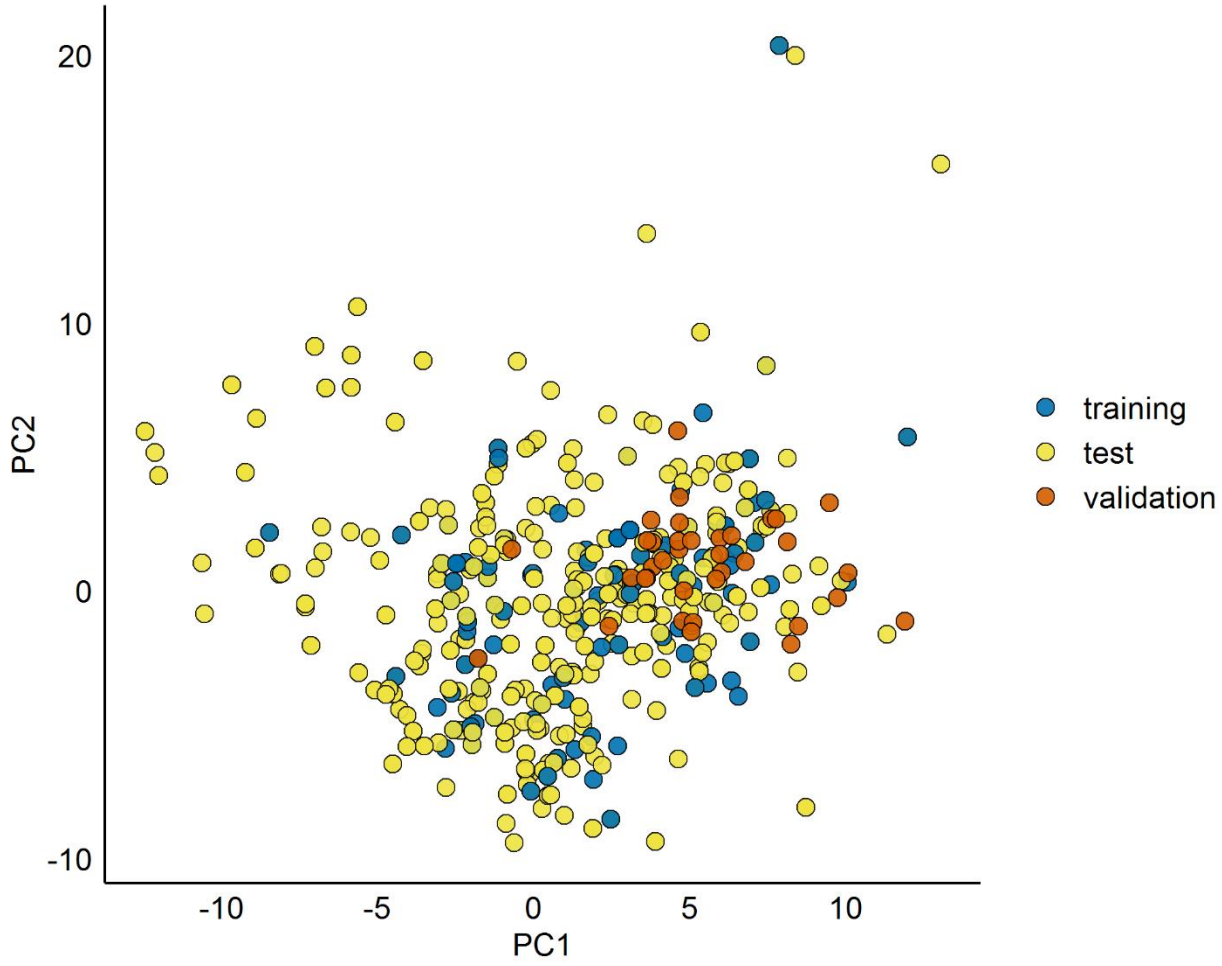
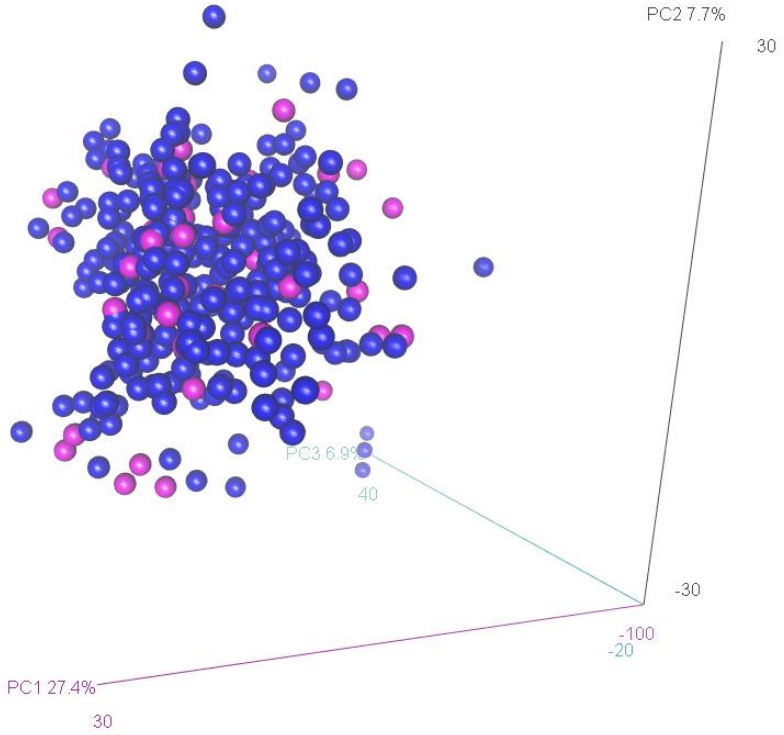


Figure S4 PCA analysis of the training dataset compounds(n = 353). Violet dots denote the compounds used for studying the solvents. For choosing the set to study solvents, 18 first principal components (described variance 70.7%) were used. For clairty the first three principal components are presented. Each dot represesnts one compound.



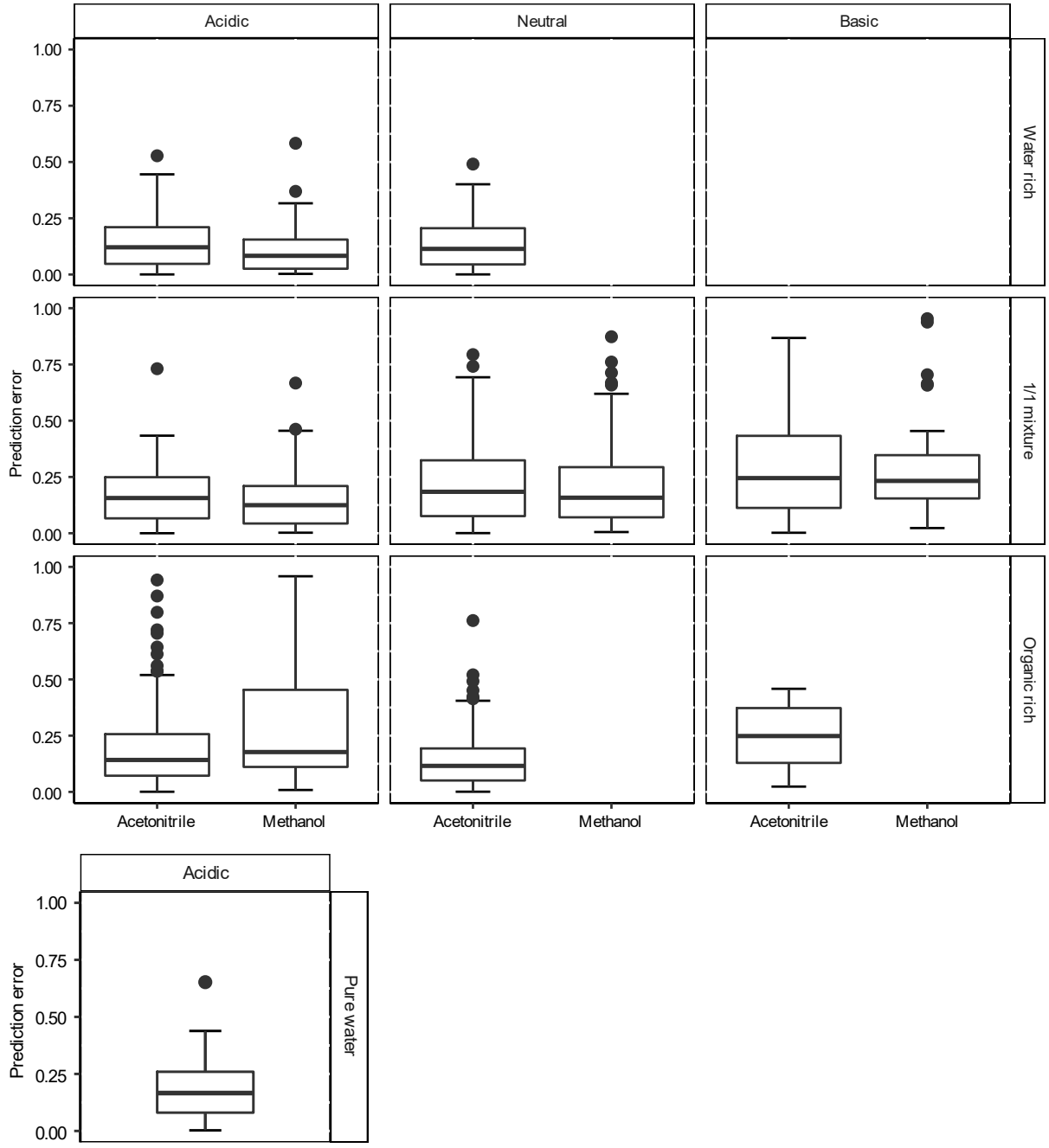
Boxplots

For the boxplots the lower line presents the 1st quartile, the line in the middle the 2nd quartile (median) and the higher line the 3rd quartile. Whiskers are found according to the formula:

$$upper\ whisker = \min(\max(x)\ , Q_3 + 1.5\ \cdot IQR) \qquad \text{Eq 5}$$
$$lower\ whisker = \max(\min(x), Q_1 - 1.5 \cdot IQR\ )$$

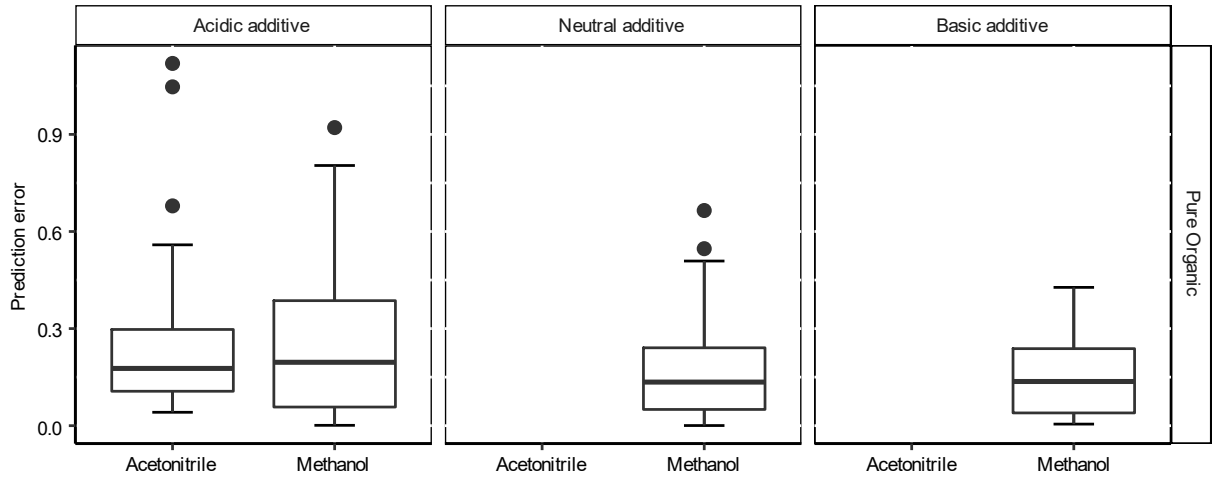
where $IQR = Q_3 - Q_1$ and Q_1 represents 1st quartile and Q_3 represents 3rd quartile.

Figure S5 Comparison of prediction errors ionization efficiencies between acetonitrile and methanol containing solvents in ESI positive mode. Results are divided into groups by water phase pH and organic modifier content. Comparison based on intersection of compounds measured in methanol as well as in acetonitrile. The compared results are measured on one instrument. Each datapoint corresponds to one compound solvent combination. Dots represent outliers.



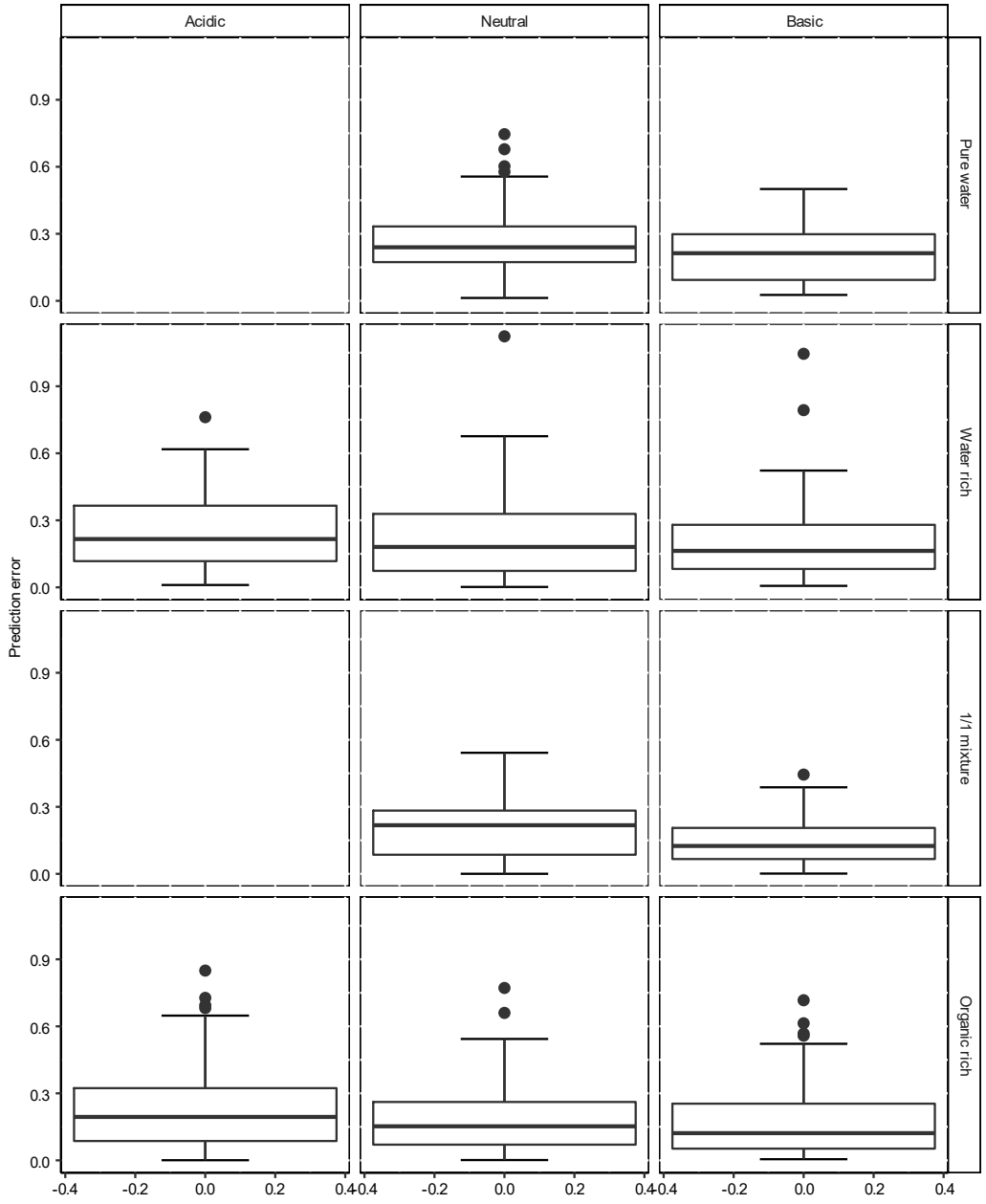
Acidic: pH < 5. **Neutral:** 5 ≤ pH < 8. **Basic:** pH ≥ 8. **Pure water:** organic modifier percentage = 0%. **Water rich:** 0% < organic modifier percentage < 40%. **1/1 mixture:** 40% ≤ organic modifier percentage < 60%. **Organic rich:** 60% ≤ organic modifier percentage < 100%.

Figure S6 Comparison of the prediction error of ionization efficiencies between neat acetonitrile and methanol in ESI positive mode. Divided into groups by pH adjusting additive type. Comparison is based on intersection of compounds measured in methanol as well as in acetonitrile. The compared results are measured on one instrument. Every datapoint corresponds to one compound. Dots represent outliers.



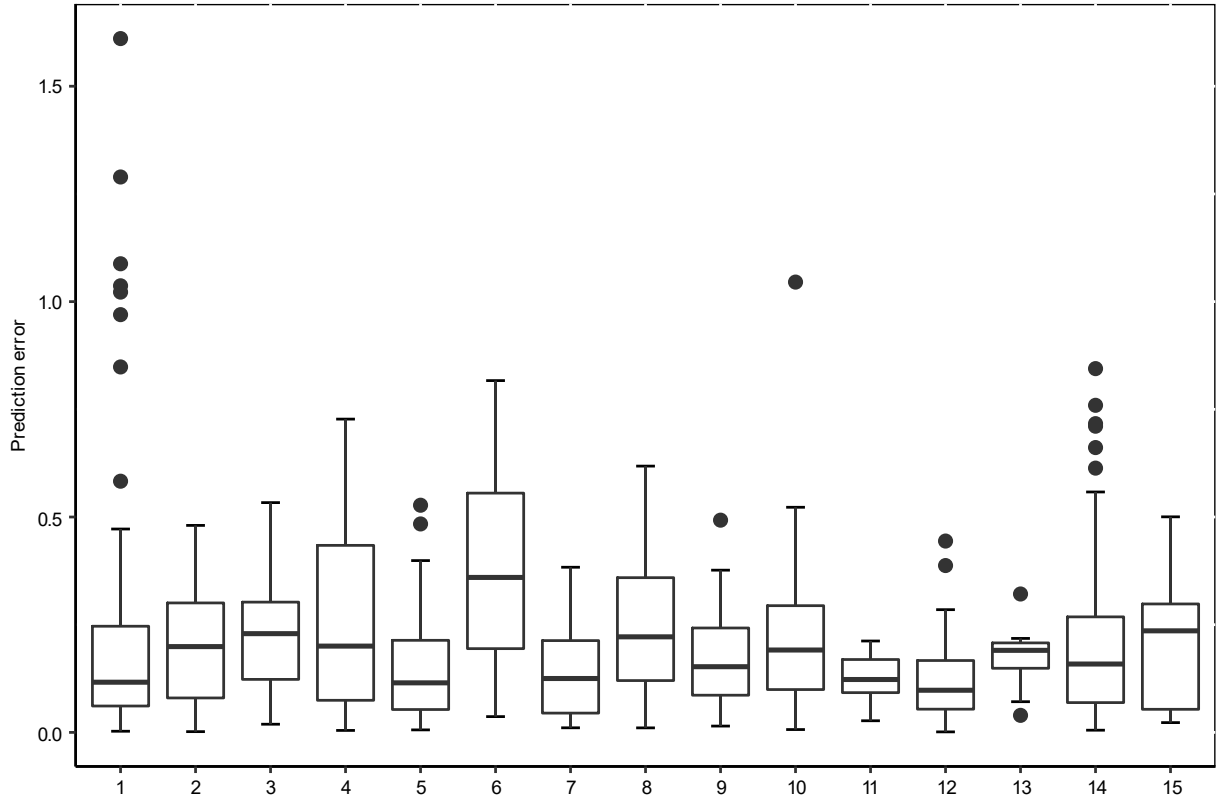
Acidic additive: formic acid, trifluoroacetic acid, oxalic acid, **Neutral additive:** ammonium acetate, ammonium formate, **Basic additive:** ammonia.

Figure S7 Comparison of the prediction error of ionization efficiencies in acetonitrile containing solvents in ESI negative mode. Results are divided into groups by water phase pH and organic modifier content. Comparison based on intersection of compounds measured in all pH groups. The compared results are measured on one instrument. Every datapoint corresponds to one compound. Dots represent outliers.



Acidic: pH < 5. **Neutral:** 5 ≤ pH < 8. **Basic:** pH ≥ 8. **Pure water:** organic modifier percentage = 0%. **Water rich:** 0% < organic modifier percentage < 40%. **1/1 mixture:** 40% ≤ organic modifier percentage < 60%. **Organic rich:** 60% ≤ organic modifier percentage < 100%.

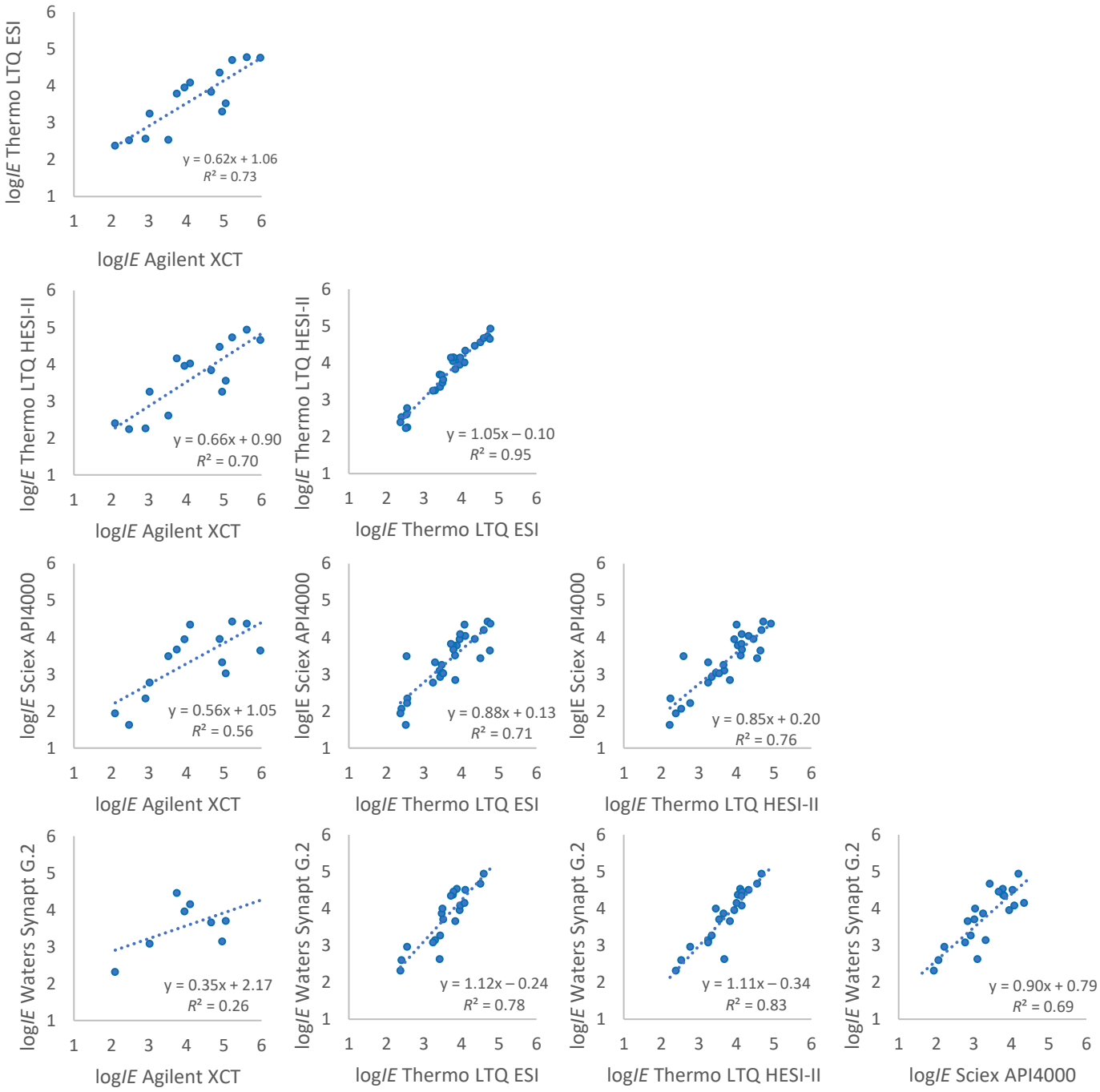
Figure S8 Comparison of the prediction error of ionization efficiency between different solvents in ESI negative mode. Compared for the intersection of compounds in studied eluent compositions. Every datapoint corresponds to one compound solvent composition combination. Dots represent outliers.



1 - acetonitrile 52 mM ammonia.
2 - acetonitrile/ water phase 20/80 4 mM ammonium acetate pH(aq) = 5.0
3 - acetonitrile/ water phase 80/20 0.2 mM ammonium acetate pH(aq) = 5.0
4 - acetonitrile/ water phase 80/20 1 mM ammonium acetate pH(aq) = 3.45
5 - acetonitrile/ water phase 80/20 1 mM ammonium acetate pH(aq) = 5.0
6 - acetonitrile/ water phase 80/20 1 mM ammonium acetate pH(aq) = 7.0
7 - acetonitrile/ water phase 80/20 1 mM ammonium acetate pH(aq) = 7.8
8 - acetonitrile/ water phase 20/80 21 mM formic acid pH(aq) = 2.78
9 - acetonitrile/ water phase 80/20 5 mM formic acid pH(aq) = 2.78
10 - acetonitrile/ water phase 20/80 41 mM ammonia pH(aq) = 10.5
11 - acetonitrile/ water phase 40/60 31 mM ammonia pH(aq) = 10.5
12 - acetonitrile/ water phase 50/50 26 mM ammonia pH(aq) = 10.5
13 - acetonitrile/ water phase 60/40 21 mM ammonia pH(aq) = 10.5
14 - acetonitrile/ water phase 80/20 10 mM ammonia pH(aq) = 10.5
15 – 52 mM ammonia pH(aq) = 10.5

Figure S9 Correlation between log*IE* values measured for set of compounds in the same eluent composition on different mass spectrometric setups. The intersection of compound-solvent cominations studied with Agilent XCT and Waters Synapt G2 is too few to study the correlations. Each dot represents one compound-solvent pair.

a.



b

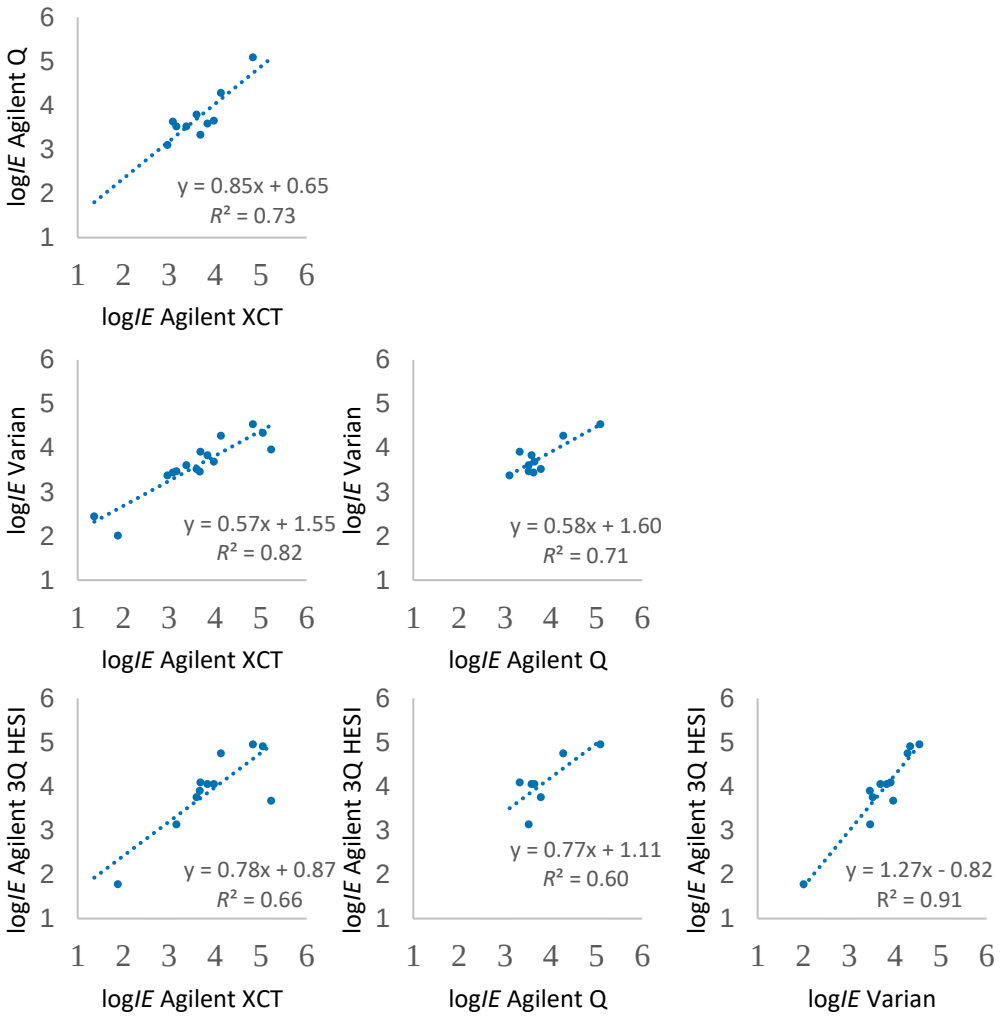

log*IE* Agilent Q
log*IE* Agilent XCT
y = 0.85x + 0.65
R^2 = 0.73
log*IE* Varian
log*IE* Agilent XCT
y = 0.57x + 1.55
R^2 = 0.82
log*IE* Varian
log*IE* Agilent Q
y = 0.58x + 1.60
R^2 = 0.71
log*IE* Agilent 3Q HESI
log*IE* Agilent XCT
y = 0.78x + 0.87
R^2 = 0.66
log*IE* Agilent 3Q HESI
log*IE* Agilent Q
y = 0.77x + 1.11
R^2 = 0.60
log*IE* Agilent 3Q HESI
log*IE* Varian
y = 1.27x - 0.82
R^2 = 0.91

Figure S10 Comparison of the prediction error of ionization efficiency values for different instruments (Table S7) in ESI positive mode. Every datapoint corresponds to one compound solvent composition combination. Dots represent outliers.

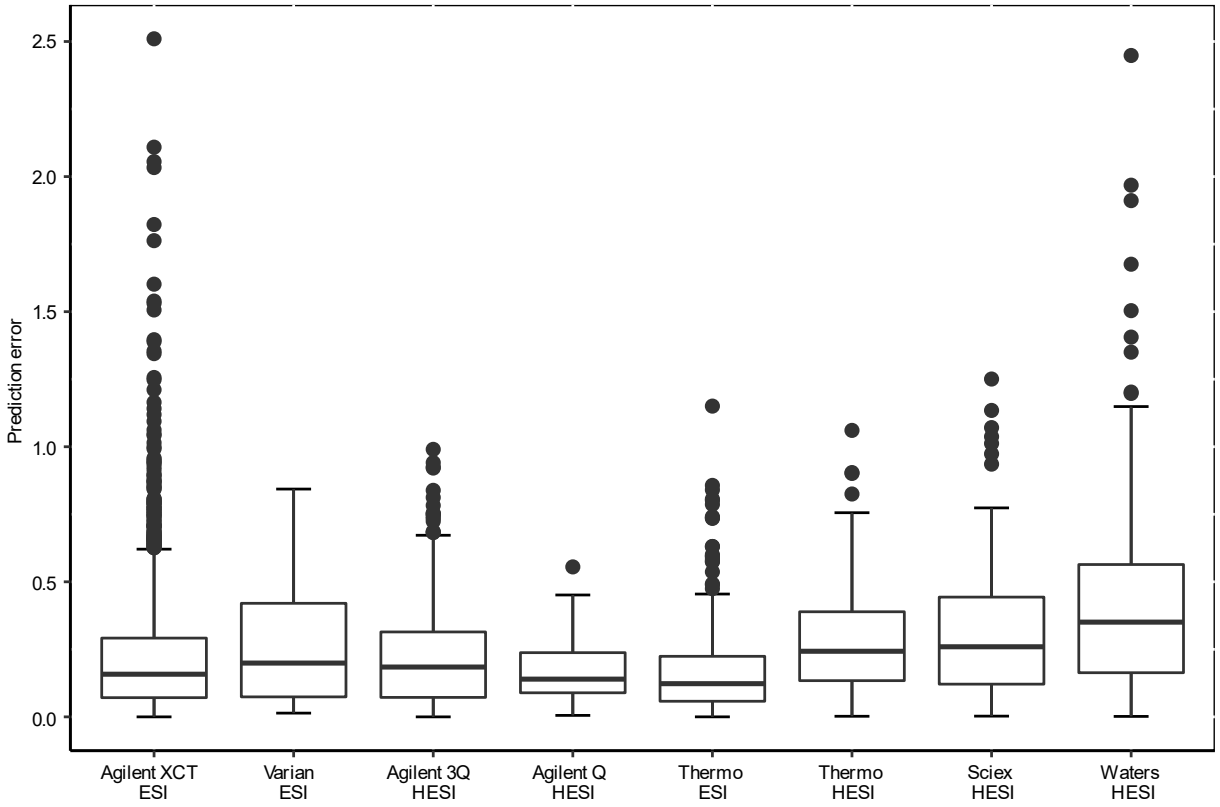


Figure S11 The predicted ionization efficiency values for the validation compounds relative to the measured values. The validation set included 35 pesticides and mycotoxines. Every datapoint corresponds to one compound solvent composition combination. All measurements have been done on Agilent 6495 triple quadrupole instrument with Jet Streem ionization source.

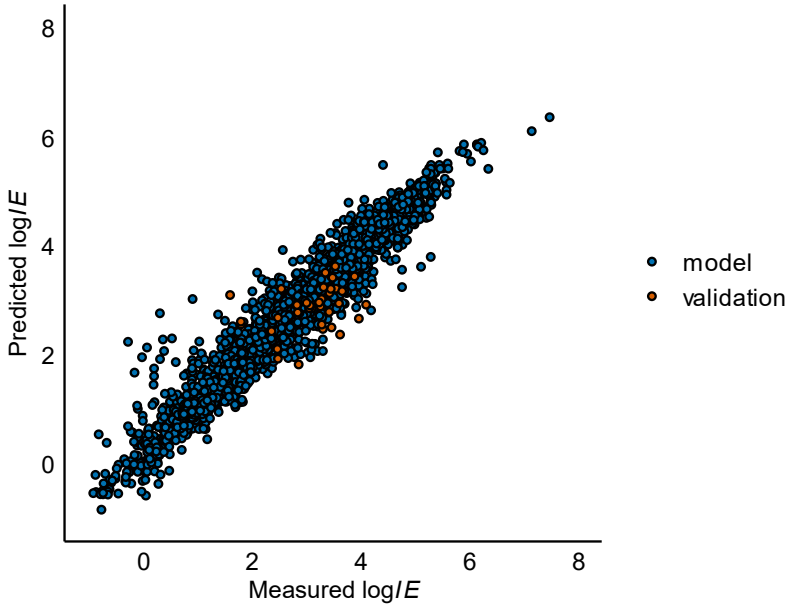


Figure S12 Comparison of predicted and spiked concentration in case of pesticides in cereal samples.

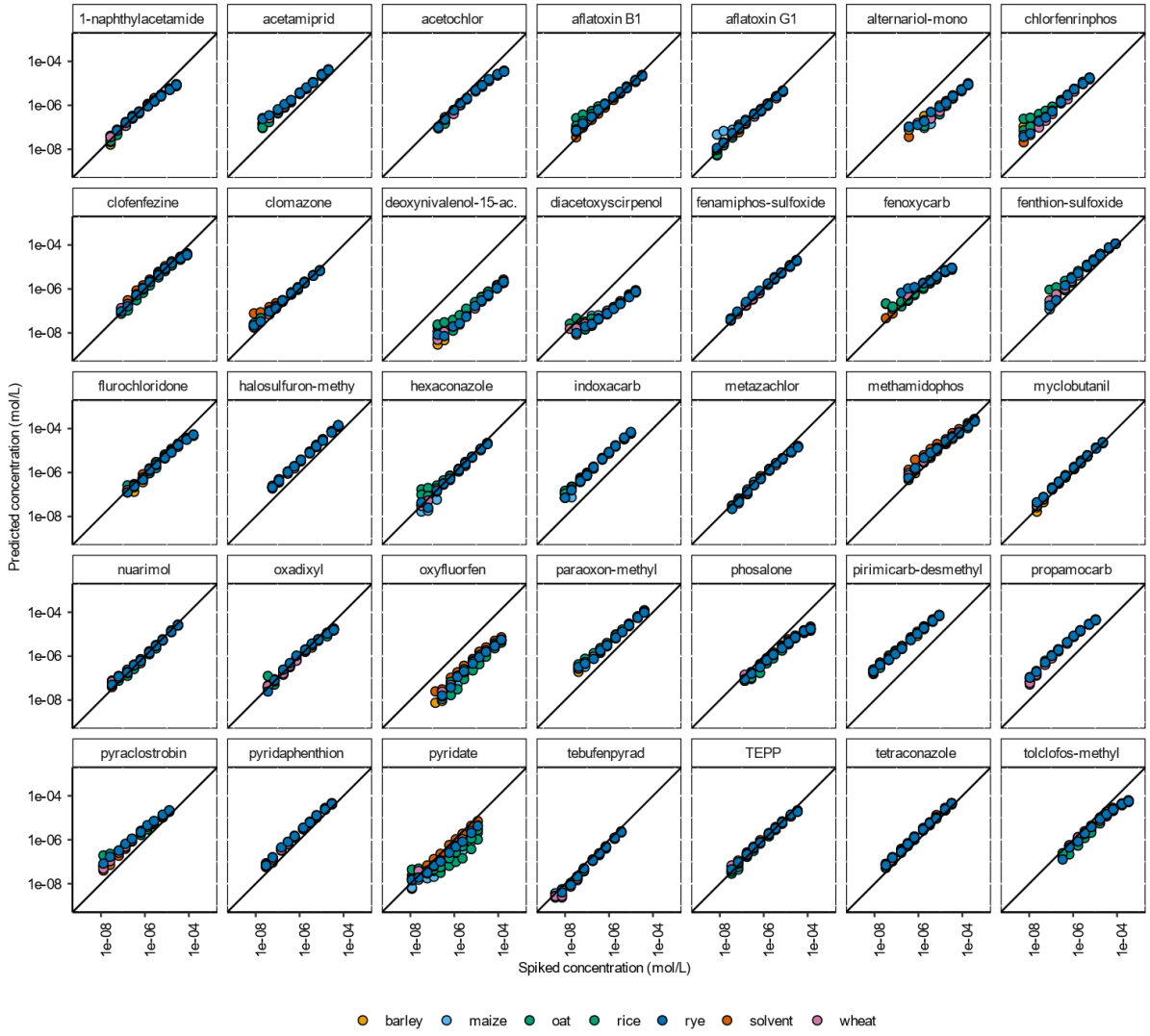


Figure S13 Comparison of the prediction error of ionization efficiencies for different ionization efficiency groups in ESI positive mode. Every datapoint corresponds to one compound-solvent combination. Dots represent outliers.

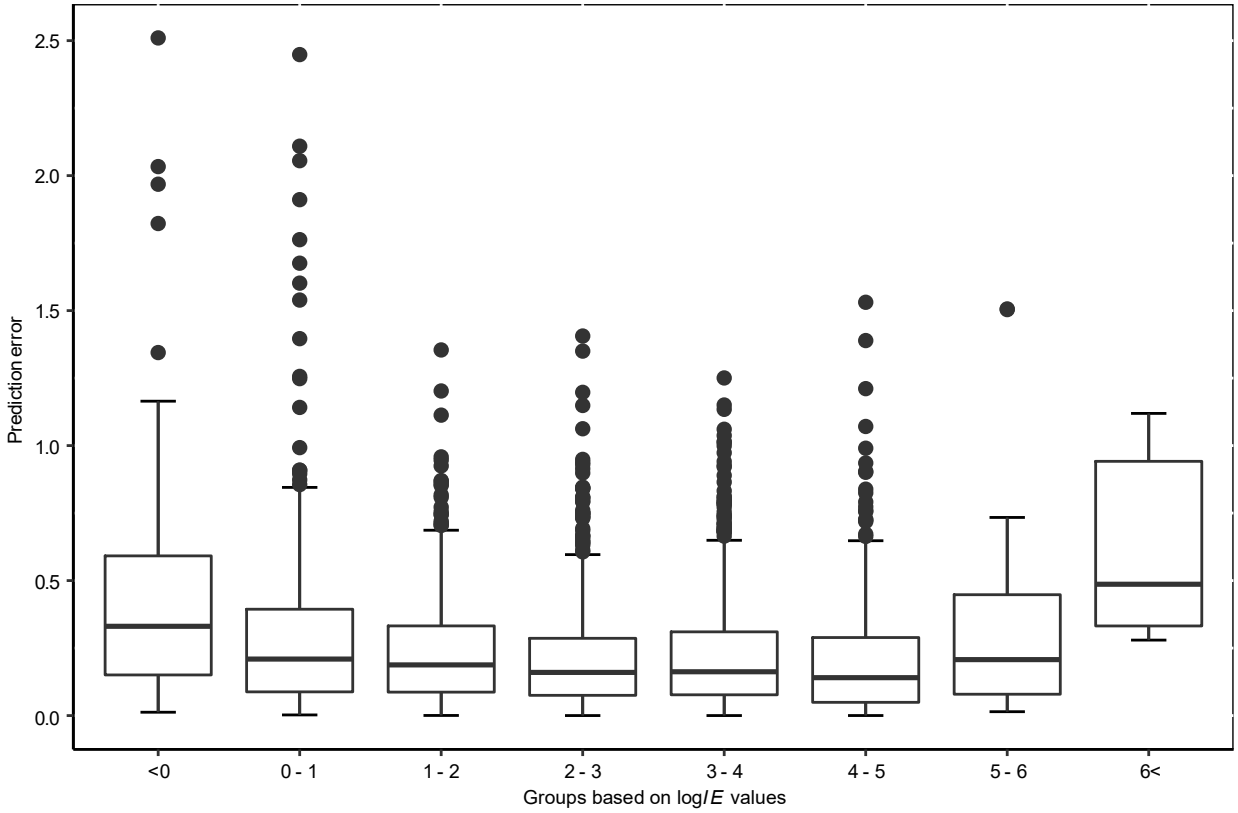
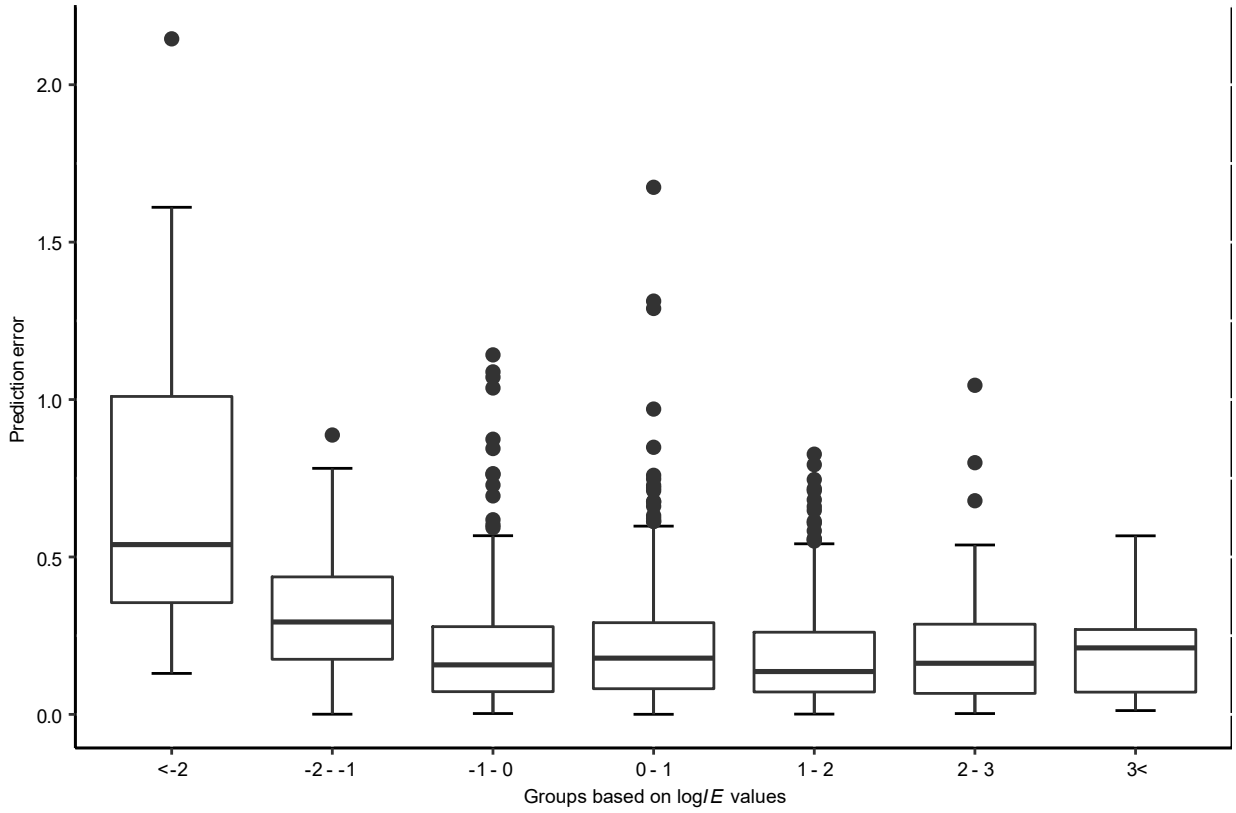


Figure S14 Comparison of the prediction error of ionization efficiencies for different ionization efficiency groups in ESI negative mode. Every datapoint corresponds to one compound-solvent combination. Dots represent outliers.



Code S1 Code used for model development

```
library(tidyverse)
library(caTools)
#input data with eluent and compound descriptors
dataset

# Splitting the dataset into the Training set and Test set
set.seed(1000)
split <-  sample.split(dataset %>% select(logIE),
                       SplitRatio = 0.8)
dataset <- dataset %>%
  mutate(SPLIT = split)

training_set <- filter(dataset,
                       SPLIT == TRUE)
test_set <- filter(dataset,
                   SPLIT == FALSE)
#Selecting only parameters that are needed for model development
training_set <- select(training_set, -SPLIT)
test_set <- select(test_set, -SPLIT)

# Random forest regression requires y as vector and features as matrix
y_train <-  as.vector(training_set$logIE)
y_test <-  as.vector(test_set$logIE)
x_train <-  as.matrix(training_set[,-1])
x_test <-  as.matrix(test_set[,-1])

# Regularized random forest regression
regressor <-  RRF(x = x_train,
                  y = y_train,
                  xtest = x_test,
                  ytest = y_test,
                  flagReg = 0, # 1 regularization applied, 0 regularization
not applied
                  ntree = 100, # number of trees
                  keep.forest=TRUE) # needed to save the optimized random
forest
# Predicting the logIE using developd regressor
training_pred <-  predict(regressor,
                          newdata = training_set,
                          predict.all = TRUE)$aggregate
test_pred <- predict(regressor,
                     newdata = test_set,
                     predict.all = TRUE)$aggregate
```

References

1. Snyder, L. R., Kirkland, J. J. & Dolan, J. W. *Introduction to Modern Liquid Chromatography*. (John Wiley & Sons, Inc., 2009).
2. Mikhail, S. Z. & Kimel, W. R. Densities and Viscosities of 1-Propanol-Water Mixtures. *J. Chem. Eng. Data* **8**, 323–328 (1963).
3. Noda, K., Ohashi, M. & Ishida, K. Viscosities and Densities at 298.15 K for Mixtures of Methanol, Acetone, and Water. *J Chem Eng Data* **27**, 326–328 (1982).
4. Rudakov, O. B., Belyaev, D. S., Khorokhordina, E. A. & Podolina, E. A. Surface tension of binary mobile phases for liquid chromatography. *Russ. J. Phys. Chem. A* **81**, 366–369 (2007).
5. Howard, K. S. & McAllister, R. A. Surface tension of acetone-water solutions up to their normal boiling points. *AIChE J.* **3**, 325–329 (1957).
6. Katz, E., Eksteen, R., Schoenmakers, P. & Miller, N. *Handbook of HPLC*. (M. Dekker, 1998).
7. Alfaro, C. M., Uwakweh, A.-O., Todd, D. A., Ehrmann, B. M. & Cech, N. B. Investigations of Analyte-Specific Response Saturation and Dynamic Range Limitations in Atmospheric Pressure Ionization Mass Spectrometry. *Anal. Chem.* **86**, 10639–10645 (2014).
8. Alymatiri, C. M., Kouskoura, M. G. & Markopoulou, C. K. Decoding the signal response of steroids in electrospray ionization mode (ESI-MS). *Anal Methods* **7**, 10433–10444 (2015).
9. Basiri, B., Murph, M. M. & Bartlett, M. G. Assessing the Interplay between the Physicochemical Parameters of Ion-Pairing Reagents and the Analyte Sequence on the Electrospray Desorption Process for Oligonucleotides. *J. Am. Soc. Mass Spectrom.* **28**, 1647–1656 (2017).
10. Beach, D. G. & Gabryelski, W. Linear and Nonlinear Regimes of Electrospray Signal Response in Analysis of Urine by Electrospray Ionization-High Field Asymmetric Waveform Ion Mobility Spectrometry-MS and Implications for Nontarget Quantification. *Anal. Chem.* **85**, 2127–2134 (2013).
11. Bedner, M. & Duewer, D. L. Dynamic Calibration Approach for Determining Catechins and Gallic Acid in Green Tea Using LC–ESI/MS. *Anal. Chem.* **83**, 6169–6176 (2011).
12. Byrdwell, W. C. Quadruple parallel mass spectrometry for analysis of vitamin D and triacylglycerols in a dietary supplement. *J. Chromatogr. A* **1320**, 48–65 (2013).
13. Caetano, S. *et al.* Exploring and modelling the responses of electrospray and atmospheric pressure chemical ionization techniques based on molecular descriptors. *Anal. Chim. Acta* **550**, 92–106 (2005).
14. Cech, N. B. & Enke, C. G. Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides. *Anal. Chem.* **72**, 2717–2723 (2000).
15. Cech, N. B., Krone, J. R. & Enke, C. G. Predicting Electrospray Response from Chromatographic Retention Time. *Anal. Chem.* **73**, 208–213 (2001).
16. Chalcraft, K. R., 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.* **81**, 2506–2515 (2009).
17. Cífková, E. *et al.* Nontargeted Quantitation of Lipid Classes Using Hydrophilic Interaction Liquid Chromatography–Electrospray Ionization Mass Spectrometry with Single Internal Standard and Response Factor Approach. *Anal. Chem.* **84**, 10064–10070 (2012).
18. Cramer, C. J., Johnson, J. L. & Kamel, A. M. Prediction of Mass Spectral Response Factors from Predicted Chemometric Data for Druglike Molecules. *J. Am. Soc. Mass Spectrom.* **28**, 278–285 (2017).

19. Dahal, U. P., Jones, J. P., Davis, J. A. & Rock, D. A. Small Molecule Quantification by Liquid Chromatography-Mass Spectrometry for Metabolites of Drugs and Drug Candidates. *Drug Metab. Dispos.* **39**, 2355–2360 (2011).
20. Ehrmann, B. M., Henriksen, T. & Cech, N. B. Relative importance of basicity in the gas phase and in solution for determining selectivity in electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.* **19**, 719–728 (2008).
21. Espinosa, M. S., Folguera, L., Magallanes, J. F. & Babay, P. A. Exploring analyte response in an ESI-MS system with different chemometric tools. *Chemom. Intell. Lab. Syst.* **146**, 120–127 (2015).
22. Gao, F. *et al.* A Refined Model for Ionization of Small Molecules in Electrospray Mass Spectrometry. *Chem. Lett.* **45**, 955–957 (2016).
23. Ghosh, B. & Jones, A. D. Dependence of negative-mode electrospray ionization response factors on mobile phase composition and molecular structure for newly-authenticated neutral acylsucrose metabolites. *The Analyst* **140**, 6522–6531 (2015).
24. Gioumouxouzis, C. I., Kouskoura, M. G. & Markopoulou, C. K. Negative electrospray ionization mode in mass spectrometry: A new perspective via modeling. *J. Chromatogr. B* **998–999**, 97–105 (2015).
25. Golubović, J., Birkemeyer, C., Protić, A., Otašević, B. & Zečević, M. Structure–response relationship in electrospray ionization-mass spectrometry of sartans by artificial neural networks. *J. Chromatogr. A* **1438**, 123–132 (2016).
26. Hatsis, P., Waters, N. J. & Argikar, U. A. Implications for Metabolite Quantification by Mass Spectrometry in the Absence of Authentic Standards. *Drug Metab. Dispos.* **45**, 492–496 (2017).
27. Henriksen, T., Juhler, R. K., Svensmark, B. & Cech, N. B. The relative influences of acidity and polarity on responsiveness of small organic molecules to analysis with negative ion electrospray ionization mass spectrometry (ESI-MS). *J. Am. Soc. Mass Spectrom.* **16**, 446–455 (2005).
28. Hermans, J., Ongay, S., Markov, V. & Bischoff, R. Physicochemical Parameters Affecting the Electrospray Ionization Efficiency of Amino Acids after Acylation. *Anal. Chem.* **89**, 9159–9166 (2017).
29. Huffman, B. A., Poltash, M. L. & Hughey, C. A. Effect of Polar Protic and Polar Aprotic Solvents on Negative-Ion Electrospray Ionization and Chromatographic Separation of Small Acidic Molecules. *Anal. Chem.* **84**, 9942–9950 (2012).
30. Kalogiouri, N. P., Aalizadeh, R. & Thomaidis, N. S. Investigating the organic and conventional production type of olive oil with target and suspect screening by LC-QTOF-MS, a novel semi-quantification method using chemical similarity and advanced chemometrics. *Anal. Bioanal. Chem.* **409**, 5413–5426 (2017).
31. Kamga, A. W., Behar, F. & Hatcher, P. G. Quantitative Analysis of Long Chain Fatty Acids Present in a Type I Kerogen Using Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: Compared with BF3/MeOH Methylation/GC-FID. *J. Am. Soc. Mass Spectrom.* **25**, 880–890 (2014).
32. Kiontke, A., Oliveira-Birkmeier, A., Opitz, A. & Birkemeyer, C. Electrospray Ionization Efficiency Is Dependent on Different Molecular Descriptors with Respect to Solvent pH and Instrumental Configuration. *PLOS ONE* **11**, e0167502 (2016).
33. Koivusalo, M., Haimi, P., Heikinheimo, L., Kostiainen, R. & Somerharju, P. Quantitative determination of phospholipid compositions by ESI-MS: effects of acyl chain length, unsaturation, and lipid concentration on instrument response. *J. Lipid Res.* **42**, 663–672 (2001).

34. Leitner, A., Emmert, J., Boerner, K. & Lindner, W. Influence of Solvent Additive Composition on Chromatographic Separation and Sodium Adduct Formation of Peptides in HPLC–ESI MS. *Chromatographia* **65**, 649–653 (2007).
35. Mandra, V. J., Kouskoura, M. G. & Markopoulou, C. K. Using the partial least squares method to model the electrospray ionization response produced by small pharmaceutical molecules in positive mode: Modelling positive electrospray ionization response. *Rapid Commun. Mass Spectrom.* **29**, 1661–1675 (2015).
36. Mehta, N. *et al.* Mass Spectrometric Quantification of N-Linked Glycans by Reference to Exogenous Standards. *J. Proteome Res.* **15**, 2969–2980 (2016).
37. Monnin, C., Ramrup, P., Daigle-Young, C. & Vuckovic, D. Improving negative liquid chromatography/electrospray ionization mass spectrometry lipidomic analysis of human plasma using acetic acid as a mobile-phase additive. *Rapid Commun. Mass Spectrom.* **32**, 201–211 (2018).
38. Nguyen, T. B., Nizkorodov, S. A., Laskin, A. & Laskin, J. An approach toward quantification of organic compounds in complex environmental samples using high-resolution electrospray ionization mass spectrometry. *Anal Methods* **5**, 72–80 (2013).
39. Pieke, E. N., 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* **975**, 30–41 (2017).
40. Raji, M. A. *et al.* Using multivariate statistical methods to model the electrospray ionization response of GXG tripeptides based on multiple physicochemical parameters. *Rapid Commun. Mass Spectrom.* **23**, 2221–2232 (2009).
41. Stavenhagen, K. *et al.* Quantitative mapping of glycoprotein micro-heterogeneity and macro-heterogeneity: an evaluation of mass spectrometry signal strengths using synthetic peptides and glycopeptides: Glycopeptide ionisation strength. *J. Mass Spectrom.* **48**, 627–639 (2013).
42. Zendong, Z., Sibat, M., Herrenknecht, C., Hess, P. & McCarron, P. Relative molar response of lipophilic marine algal toxins in liquid chromatography/electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **31**, 1453–1461 (2017).
43. Tang, W.-T., Fang, M.-F., Liu, X. & Yue, M. Simultaneous Quantitative and Qualitative Analysis of Flavonoids from Ultraviolet-B Radiation in Leaves and Roots of *Scutellaria baicalensis* Georgi Using LC-UV-ESI-Q/TOF/MS. *J. Anal. Methods Chem.* **2014**, 1–9 (2014).
44. Tu, J., Yin, Y., Xu, M., Wang, R. & Zhu, Z.-J. Absolute quantitative lipidomics reveals lipidome-wide alterations in aging brain. *Metabolomics* **14**, (2018).
45. Wu, L. *et al.* Quantitative structure–ion intensity relationship strategy to the prediction of absolute levels without authentic standards. *Anal. Chim. Acta* **794**, 67–75 (2013).
46. Yang, W.-C., Mirzaei, H., Liu, X. & Regnier, F. E. Enhancement of Amino Acid Detection and Quantification by Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **78**, 4702–4708 (2006).
47. Yang, J. *et al.* A chemical profiling strategy for semi-quantitative analysis of flavonoids in Ginkgo extracts. *J. Pharm. Biomed. Anal.* **123**, 147–154 (2016).
48. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminformatics* **8**, (2016).
49. Ghose, A. K. & Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **7**, 565–577 (1986).

50. Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics*. (Wiley-VCH, 2009).
51. Wildman, S. A. & Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **39**, 868–873 (1999).
52. Hall, L. H. & Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **35**, 1039–1045 (1995).
53. Roy, K. & Ghosh, G. QSTR with Extended Topochemical Atom Indices. 2. Fish Toxicity of Substituted Benzenes. *J. Chem. Inf. Comput. Sci.* **44**, 559–567 (2004).
54. Liu, S., Cao, C. & Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, λ. *J. Chem. Inf. Comput. Sci.* **38**, 387–394 (1998).
55. Platts, J. A., Butina, D., Abraham, M. H. & Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **39**, 835–845 (1999).
56. Pearlman, R. S. & Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **39**, 28–35 (1999).
57. Kier, L. B. & Hall, L. H. *Molecular connectivity in chemistry and drug research*. (Academic Press, 1976).
58. Randic, M. On molecular identification numbers. *J. Chem. Inf. Comput. Sci.* **24**, 164–175 (1984).
59. Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **69**, 17–20 (1947).
60. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
61. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).