Naive Bayes classification model for isotopologue detection in LC-HRMS data

Denice van Herwerden,[*,†] Jake W. O'Brien,[‡] Phil M. Choi,[¶,‡] Kevin V. Thomas,[‡] Peter J. Schoenmakers,[†] and Saer Samanipour[*,†,‡,§]

†*Van 't Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Amsterdam*

‡*Queensland Alliance for Environmental Health Sciences (QAEHS), The University of Queensland, Australia*

¶*Water Unit, Health Protection Branch, Prevention Division, Queensland Department of Health, Australia*

§*Norwegian Institute for Water Research (NIVA), Oslo, Norway*

E-mail: d.vanherwerden@uva.nl; s.samanipour@uva.nl

Abstract

Isotopologue identification or removal is a necessary step to reduce the number of features that need to be identified in samples analyzed with non-targeted analysis. Currently available approaches rely on either predicted isotopic patterns or an arbitrary mass tolerance, requiring information on the molecular formula or instrumental error, respectively. Therefore, a Naive Bayes isotopologue classification model was developed that does not depend on any thresholds or molecular formula information. This classification model uses elemental mass defects of six elemental ratios and can successfully identify isotopologues in both theoretical isotopic patterns and wastewater influent samples, outperforming one of the most commonly used approaches (i.e., 1.0033 Da mass difference method - CAMERA).

Introduction

Non-target analysis (NTA) in combination with liquid chromatography high-resolution mass spectrometry (LC-HRMS) is a comprehensive approach for the characterization of unknown chemicals in complex sample matrices, originating from, for example, environmental or biological backgrounds.[1–6] These samples can contain thousands of structurally known and unknown chemicals. To identify these chemicals, the raw data files need to be processed to extract and group information that belongs to unique chemical constituents (i.e., parent, isotopologue, adduct, and (in-source) fragment ions).[1] During this step, one approach to reduce the number of individual features requiring identification is the detection or removal of isotopologues (i.e., heavier versions of the same monoisotopic peak).

For LC-HRMS data, two main approaches have been used to detect isotopologues.[7,8] The first strategy relies on a predicted molecular formula, which can be translated to a predicted isotopic pattern.[7,9] The main shortcoming of this approach is the difficulties associated with accurate and reliable molecular formula prediction for unknown chemical constituents. The wrong molecular formula could be assigned to a feature either due to instrumental error or absence of a chemical constituent in a database. These wrongly assigned molecular formulas could lead to identifying the potential isotopologues of a feature with the wrong isotopic pattern, resulting in higher false positive and false negative identification rates.

On the other hand, a theoretical mass difference of n $\times$ 1.0033 Da (i.e., CAMERA) has been used.[8,10] Here n equals the depth of the isotopopologue mass. For example, an isotopologue mass depth of four corresponds to the mass range of the monoisotopic peak plus three isotopologues. This approach, even though elegant given that it does not require information on the molecular formula, does require an arbitrary mass tolerance as input. This means that the mass tolerance changes, depending on the instrument used, and needs to be correctly provided by the user.

In this manuscript, an isotopologue classification model is proposed that requires no prior knowledge of the molecular composition or arbitrary tolerances. The Naive Bayes classification model was generated using elemental mass defects, for which the potential in isotopologue detection was explored. For performance evaluation of the classification model, a comparison was made with an "in-house" developed mass difference method. This comparison was performed for both theoretical isotopic patterns and wastewater influent samples.

Experimental Section

LC-HRMS Analysis

The fourty-four Wastewater influent and three quality control samples were analyzed with LC-HRMS. Briefly, samples were collected over a time window of 24 hours, using on-site autosamplers set to use the optimized conditions described by Ort et al.[11] These samples were filtered, spiked with 10 ng L^{-1} of 19 labeled internal standards, and stored frozen until analysis. For analysis, 10 μL of the sample was injected on a biphenyl column at 45°C and separated using a 10-minute gradient from 5 to 100% methanol with 0.1% formic acid. The eluent was analyzed using a QToF in positive ion mode with a mass range of 50 to 600 Da and collision energy of 10 eV. Further details on the analysis are provided elsewhere.[12]

Data Processing

The raw data files were converted to mzXML file format, using MSConvertGUI (64-bit, ProteoWizard[13]). Feature lists were generated with the self adjusting feature detection (SAFD) algorithm, using the following settings: 10 000 maximum number of iterations, a minimum intensity of 500, resolution of 20 000, 0.02 m/z minimum window size in the mass domain, 0.75 minimum regression coefficient, a maximum signal increment of 5, a signal to noise ratio

of 2, and a minimum and maximum peak width in the time domain of 3 and 200 s, respectively.[12] These feature lists were used for the performance evaluation of the classification model on real samples.

Theoretical Isotopic Patterns

The isotopic patterns used for setting up the probabilistic isotopologue classification model were calculated for 737 594 chemicals from the DDS-TOX database.[14] These chemicals consist of a curated list of compounds relevant to environmental and human health. The isotopic patterns were obtained using pyOpenMS[9] (v2.6.0), combining both the isotopic masses from the fine[15,16] and coarse[9] isotope pattern generator. The fine isotope pattern generator calculates the hyperfine isotopic pattern that is obtained when the mass defect of isotopes in taken into account.[9] This mass defect equals the difference between the actual mass of an atom and the sum of the building blocks (e.g., neutrons) the atom is comprised of. From this method, isotopologues with a maximum unexplained probability of 0.01% was used. On the other hand, the coarse isotope pattern generator calculates the unit mass isotopic patterns, using the summed probability for each isotopologue peak, ignoring the hyperfine structures. For this, a maximum isotopic tree depth was required that corresponds to one plus the maximum number of isotopes that could be present in a single molecule.[16] Considering the fact that an increasing number of isotopes within a molecule results in a lower occurrence probability (i.e., intensity), a maximum isotopic tree depth of 6 was chosen.

The full isotopic pattern for a compound was comprised of the fine and coarse isotopic patterns, excluding duplicate isotopologues from the coarse isotopic pattern that had a mass difference of $\leq$ than 0.003 Da with any of the other isotopologues, which is the typical mass error observed in LC-HRMS experiments.[17] In this manuscript, a monoisotopic parent ion with one of its isotopologues is referred to as an mono-iso pair. For example, if a monoiso-

topic parent ion has 5 theoretical isotopologues, 5 mono-iso pairs are obtained. In total, 2 691 244 mono-iso pairs were generated, which were employed for training (85% of the mono-iso pairs) and testing (15% of the mono-iso pairs) of the probabilistic isotopologue classification model (available on figshare).[18]

Elemental Ratio Calculations

To construct the probabilistic isotopologue classification model, elemental mass defects (*EMD*s) were used. The assumption here is that the monoisotopic and isotopologue mass have the same *EMD* because they have the same molecular structure with the isotopologue having one or more of its atoms being replaced with heavier versions (i.e., isotopes) of the same elements. To calculate the *EMD* for both the monoisotopic and isotopologue mass, the elemental mass (*EM*) needs to be calculated according to equation 1. Here, the ion_{mass} can either be the monoisotopic or the isotopologue mass and the er_{mass} (i.e., elemental ratio mass) depends on the elemental ratio used. For the classification model the elemental ratios CO, CCl, CN, CS, CF, and CH were used, which have an er_{mass} of 27.995, 46.969, 26.003, 43.972, 30.998, and 13.008, respectively. These values are the sum of the elemental masses of each element for a single elemental ratio. For example, the er_{mass} of CO equals the monoisotopic mass of a carbon atom plus that of an oxygen atom (i.e., 12.000 + 15.995 = 27.995). The selected elemental ratios were chosen based on both the frequency they were encountered in the DDS-Tox database (Table S1) and the fact that only 0.007% of the database entries contain none of the selected elements.

After the *EM* is calculated, the *EMD* for the monoisotopic and isotopologue mass can be obtained according to equation 2 (i.e., EMD_{mono} and EMD_{iso}, respectively). These *EMD* values are used to calculate the delta *EMD* (*dEMD*) for an mono-iso pair (Equation 3). It is important to note that the EMD_{mono} should always be subtracted from the EMD_{iso} and not

vice versa when using the probabilistic isotopologue classification model described in this paper. An example case for calculating the *dEMD* value can be found in figure 1C. The full set of isotopologue and monoisotopic *EMD* values for the DDS-Tox database can be found on figshare.[18]

$$EM = ion_{mass} \times \frac{rounded\, er_{mass}}{exact\, er_{mass}} \tag{1}$$

$$EMD = roundedEM - exactEM \tag{2}$$

$$dEMD = EMD_{iso} - EMD_{mono} \tag{3}$$

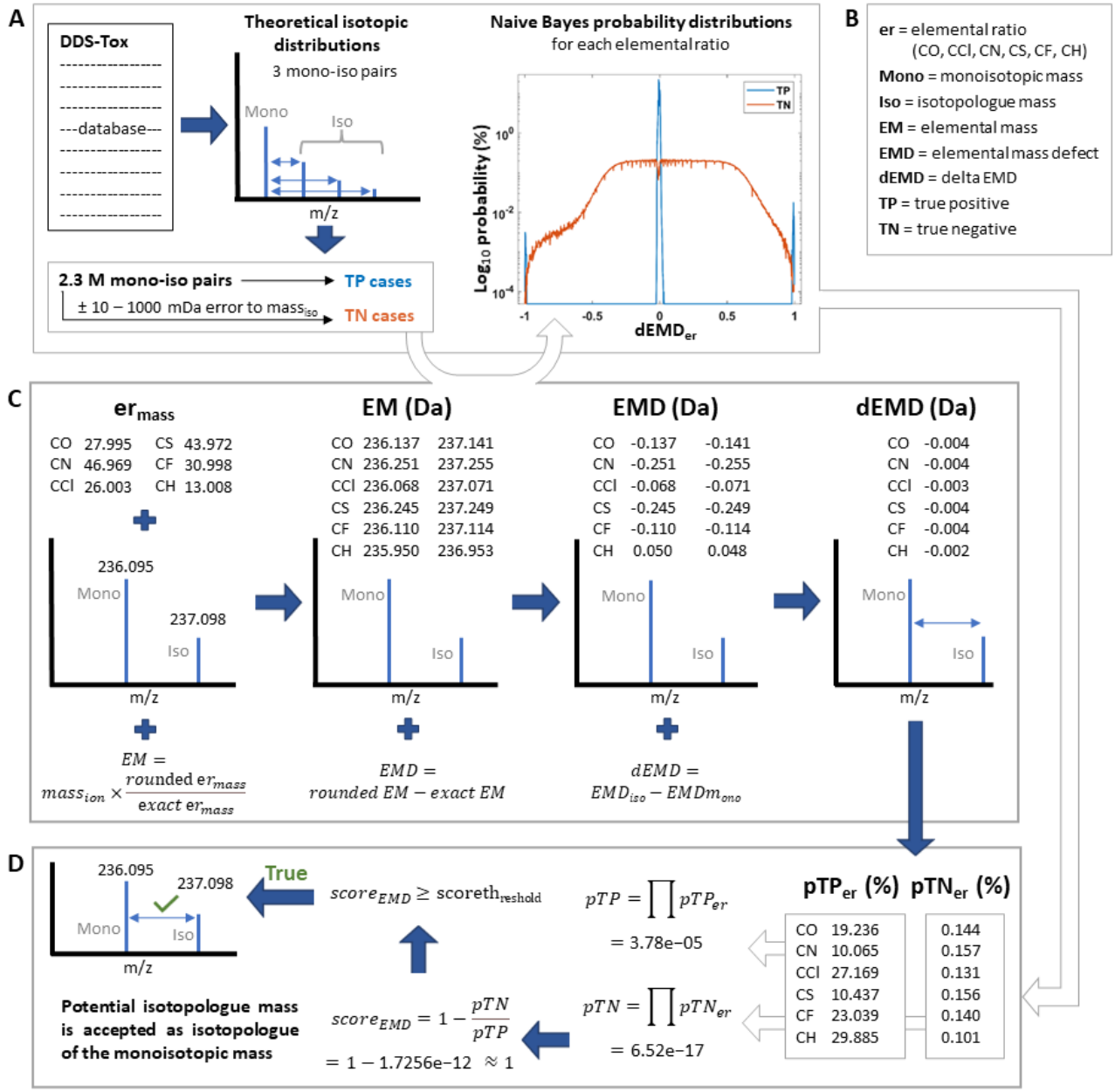

Figure 1: Section **A** shows the Workflow for the construction of the Naive Bayes isotopologue classification model, which requires calculations of the *dEMD* values (section **C**) for the mono-iso pairs. The workflow for the use of the classification model for the example mono-iso pair in **C** is shown in section **D**. Finally, **B** contains a list of abbreviations.

EMD Probability Distributions

To generate the EMD probability distributions for the classification model, both true positive (TP) and true negative (TN) mono-iso pairs were required (Figure 1A). The mono-iso

pairs in the training set were used as the true positive cases and true negative cases were generated based on the mono-iso pairs from the training set with a randomly added mass error between 0.01 to 1 Da to the isotopologue mass. For all mono-iso pairs in the TP and TN training set, the *dEMD*s were calculated for the selected elemental ratios (Equation 3). These *dEMD* values were used to construct the TP and TN probability distributions for each of the six elemental ratios. To build these probability distributions, the generated *dEMD* values were binned, using a range between -1 and 1 Da with a 0.002 Da step size. For each *dEMD* bin, the number of occurrences plus one was used. This prevented that a *dEMD* range could have a probability equal to zero, in case no occurrences for that specific *dEMD* were found in the training set. Finally, the probability distributions were calculated by dividing the occurrence distribution values by the total number of occurrences.

Naive Bayes Classification

Naive Bayes classification was used to develop a probabilistic isotopologue detection model, using the TP and TN *dEMD* probability distributions obtained for the selected elemental ratios (i.e. CO, CCl, CN, CS, CF, and CH). To calculate the posterior probabilities (i.e., P(A|B)) for classifying a potential mono-iso pair as TP or TN, Bayes theorem is used (Equation 4).[19] Here, P(A) is the probability of an mono-iso pair being TP or TN, P(B) is the occurrence likelihood for a specific *dEMD* value, P(B|A) is the probability for a *dEMD* value in case of A, and n equals the number of elemental ratios used, which would be six for our model (i.e., CO, CN, CCl, CS, CF, and CH).

$$P(A|B) = \prod_{i=1}^{n} \frac{P(B|A)_i \times P(A)_i}{P(B)_i} \tag{4}$$

Since P(B) is a marginal probability (i.e., constant probability normalizing factor), equa-

tion 4 can be rewritten to equation 5. Additionally, a uniform distribution is assumed for the prior P(A), further reducing the formula to equation 6.

$$P(A|B) \propto \prod_{i=1}^{n} P(B|A)_i \times P(A)_i \tag{5}$$

$$P(A|B) \propto \prod_{i=1}^{n} P(B|A)_i \tag{6}$$

Lastly, for the classification of the potential mono-iso pair, the TP and TN probabilities are obtained using equation 6. These probabilities were converted to probability percentages (i.e., on a scale of 0 to 100). Due to the wide range of values that can be obtained for the TP and TN probabilities, a $score_{EMD}$ is used instead for the evaluation (Equation 7). Here *P(TP)* and *P(TN)* equal the true positive and true negative probabilities, respectively. This $score_{EMD}$ ranges between 1 and minus infinity. In case the potential mono-iso pair has a $score_{EMD}$ above a set threshold, the potential isotopologue is classified as a correct isotopologue of the monoisotopic ion. An example for the calculation of the $score_{EMD}$ can be found in figure 1D

$$score_{EMD} = 1 - \frac{P(TN)}{P(TP)} \tag{7}$$

Performance Assessment

For the performance assessment, the test set was used. In this instance, TN cases were also generated based on the mono-iso pairs from the test set with a random mass error added to the isotopologue mass of 0.01 to 1 Da. For both the TP and TN cases, the $score_{EMD}$s

were calculated (Equation 7). To select a suitable $score_{EMD}$ cut-off value and assess the performance of the classification model, the TP and false positive (FP) rates were calculated for a range of $score_{EMD}$ values. The $scores_{EMD}$ fromm 0.7 to 1 Da with a step size of 0.002 Da were employed to calculate the TP_r and FP_r (Equation 8 and 9, respectively). Here, the TPs equal the number of cases from the test set that were correctly classified as an isotopologue, FNs are the number of cases that were incorrectly classified as not an isotopologue, TNs are cases that were correctly classified as not an isotopologue, and FPs are the that were wrongly classified as isotopologues.

$$TP_r = \frac{TP}{TP + FN} * 100 \tag{8}$$

$$FP_r = \frac{FP}{TN + FP} * 100 \tag{9}$$

Mass Difference Method

The mass difference method is a commonly used approach for automated isotopologue detection in LC-HRMS data. This method has already been implemented in different open access algorithms such as CAMERA and MZmine.[8,10] Here, an "in-house" developed mass difference strategy was employed to benchmark our classification model against. For the mass difference method, to asses if a signal is an isotopologue of a monoisotopic peak, first the mass difference between the signal and monoisotopic ion was calculated. Then, the residue of the division of the mass difference by 1.0033 Da is obtained. For example, if the mass difference is 2.0081 Da, the residue would be 0.0015 Da. In case the residue is lower than the specified mass tolerance, the signal is accepted as an isotopologue of the monoisotopic

mass. For the mass difference method, when dealing with the training set a mass tolerance of ± 0.0001 Da was used based on the assumption that the theoretical isotopologues do not contain any mass error. On the other hand for the wastewater samples, this mass tolerance was increased to ± 0.01 Da to better reflect the inherent mass error in such data caused by background signal and instrumental fluctuations.

Isotopologue Detection Performance for Wastewater Samples

To test the isotopologue classification model on real samples, the isotopologue detection performance was evaluated for the feature lists obtained from forty-four wastewater influent samples and three quality control samples. Additionally, a reference compound list comprised of forty-five chemicals was used, containing the monoisotopic masses (i.e., protonated molecular mass), retention times, and parent isotopologue distributions (Table S3). The isotopologue distributions for these chemicals were obtained from the isotope pattern preview tool in MZmine2 (v2.53), using the protonated molecular formula, a minimum intensity of 0.01%, a merge width of 0.0001 Da, and a charge of 1,which showed to cover an isotopologue mass depth of six.[10]

The presence of a reference compound was confirmed based on the reference retention time ± 0.1 minutes and the monoisotopic parent mass with a mass tolerance of 0.01 Da. When a reference compound monoisotopic parent mass was present, all features within a time range of ± 0.1 minutes were extracted. If a feature's mass was higher than the monoisotopic mass and lower than the monoisotopic mass plus 1.0033 × 6 (i.e., isotopologue mass depth of six), it was evaluated as a potential isotopologue with both the classification model and the mass difference method. When a model correctly identifies an isotopologue according to the reference parent isotopologue distribution, it is considered a TP case. Whereas the FP cases are incorrectly identified isotopoloues and the FN cases are the TP cases that were not

detected by a model. With these cases, the TP_r and FD_r were calculated for the classification model and mass difference method (Equation 8 and 10, respectively), which were used to compare the two isotopologue identification methods.

$$FD_r = \frac{FP}{TP + FP} * 100 \tag{10}$$

Calculations and Code Availability

All calculations were performed using a personal computer running Windows 10 Education with 12 cores and 32 GB of memory. For obtaining the theoretical isotopologues of the DDS-Tox database Python (v3.9.4) was used and for calculations related to the classification model Julia (v1.6.0) was used. The mzXML files were imported in julia using the MS_Import package, which is available at https://bitbucket.org/SSamanipour/ms_import.jl/src/master/. The code for the probabilistic isotopologue classification model is available at https://bitbucket.org/Denice_van_
This package includes both the probabilistic isotopologue classification model and functions to use the model with feature lists obtained either from SAFD[12] or other algorithms. The code for SAFD is available at https://bitbucket.org/SSamanipour/safd.jl/src/master/.

Results and Discussion

Exploring the EMD probability distributions

Calculating the *EMD* values for the theoretical isotopologues showed that the *EMD* values for the monoisotopic and isotopologue masses were similar. Figure 2 shows the *EMD* values

for the theoretical isotopic distribution of carbamazepine. In this example, a minimum and maximum absolute difference in *EMD*(i.e., *dEMD*) of 0.003 and 0.020 Da were found, respectively. Additionally, an increase in *dEMD* between the EMD_{mono} and EMD_{iso} was observed for isotopologues with a higher isotopologue mass depth. Even though the elements S and F are not present in the molecular formula of carbamazepine, a similar *EMD* trend is observed as for the elements O, N, CL, and H. On the other hand, figure S1 and S2 show that the presence of other elements (e.g., Br and P) in the molecular formula also do not influence the *EMD* values.

Overall, similar trends were observed for all theoretical isotopologue distributions with EMD values ranging from -0.5 to 0.5 Da for all six elemental ratios. To evaluate this trend, the Pearson correlation coefficients between the EMD_{mono} and EMD_{iso} values were obtained.[20] These coefficients were calculated separately for each elemental ratio and isotopologue mass depth of 1 till 6 (Table S2). The highest correlation of 1.00 was found for the elemental ratio CN with an isotopologue mass depth of 1 and the lowest value was 0.86 for both the elemental ratios CCl and CS with an isotopologue mass depth of 5 (Figure S3 and S4, respectively). Overall, the Pearson correlation coefficient decreases with a higher isotopologue mass depth except for an isotopologue mass depth of 6. It is expected that this was due to a relatively low number of mono-iso pairs with a depth of 6 (Table S2). These results showed that similar EMD values for mono-iso pairs were obtained throughout the theoretical dataset.

After calculating all *dEMD* values for the mono-iso pairs of both the TP and TN cases, the TP and TN probability percentage distributions were obtained for the selected elemental ratios (Figure 3). For the TP probability distributions, there were 2 regions for which the TP probabilities were higher than the TN probabilities. The first region being around a *dEMD* of 0, which is in accordance with the hypothesis that the monoisotopic and isotopologue mass

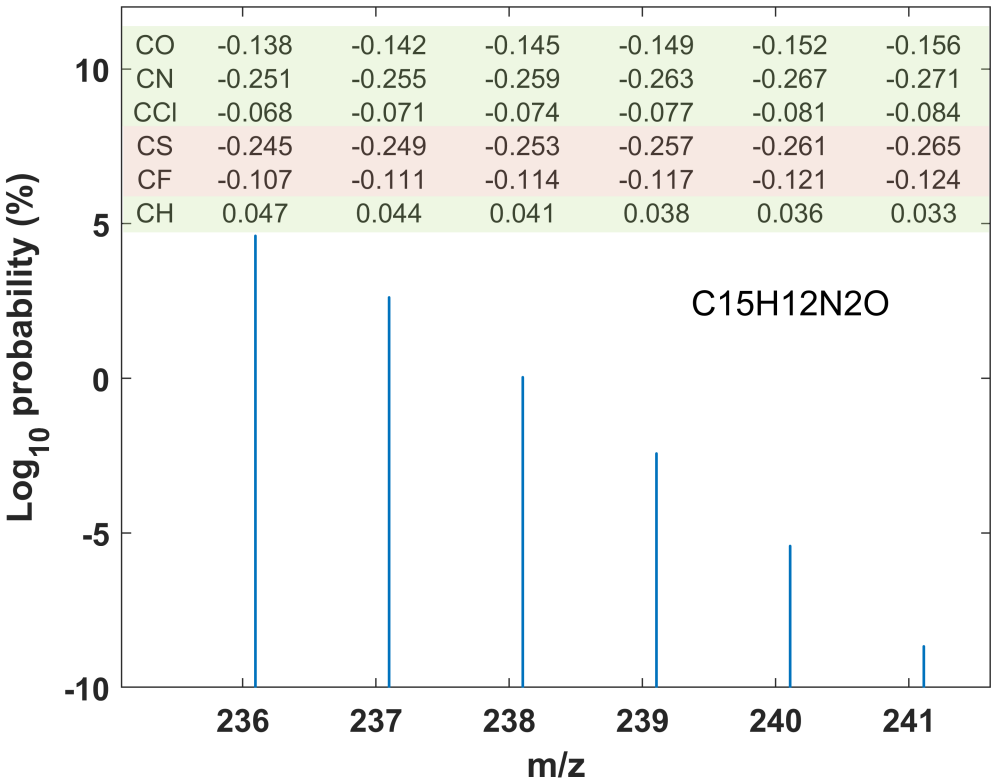

Figure 2: Isotopic distribution of carbamazepine with the corresponding $\log_{10}$ probability percentages. For the monoisotopic (236.095 Da) and each isotopologue peak (237.098, 238.102, 239.105, 240.108, and 241.112 Da), the *EMD* values are shown above in Da for the elemental ratios CO, CN, CCl, CS, CF, and CH. Additionally, the elemental ratios that are present in the molecule are marked in green and the ones that are not are marked in red.

of the same compound obtain similar *EMD* values. As for the second region, *dEMD* values close to 1 and -1 Da were found. For the TN probability distributions, a small decrease in probability was observed around a *dEMD* of 0 Da, which was caused by the minimum added mass error to the isotopologue mass of the TN mono-iso pairs (i.e., 0.01 Da). Overall, these plots showed that the *dEMD* could be used to differentiate between isotopologue and non-isotopologue masses.

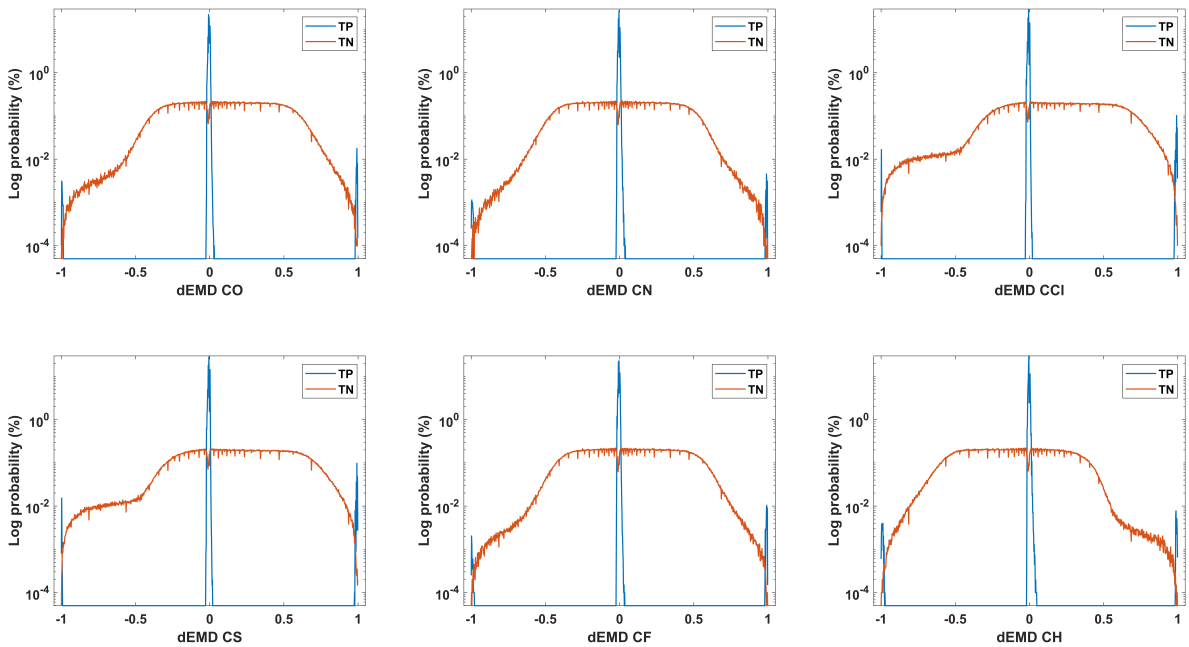

Figure 3: TP and TN probability distributions for the *dEMD* values for the selected elemental ratios CN, CCl, CO, CS, CF, and CH.

Classification Model Performance

A receiver operator curve was generated for selection of the $score_{EMD}$ threshold. This curve showed the TP_r versus the TN rate for $scores_{EMD}$ between 0.7 and 1 (Figure S5). Based on this plot a $score_{EMD}$ threshold of 0.9997 was selected. This corresponded with a TP_r and FP_r of 99.0 and 1.8%, respectively.

Comparison with existing method

To evaluate the performance of the classification model with that of the existing mass difference method, the performance for the in-house mass difference method was evaluated for a mass tolerance of 0.0001 Da. The mass tolerance was selected based on the assumption that there is no error present in the theoretical mono-iso pairs and the full receiver operator curve can be found in sectionS4. For a mass tolerance of 0.0001 Da, a TP_r and FP_r of 16.2 and 0.02% was found, respectively. Compared to the results of the classification model (i.e., TP_r of 99.0% and FP_r of 1.8%), both methods performed well with regard to the FP_r (i.e., $\leq$ 5%). However, the classification model outperformed the mass difference method for the TP_r.

Model Implementation for Real Samples

To evaluate the model performance for real samples, isotopologue detection was performed for forty-four wastewater influent and three quality control samples. A total of 391 features were evaluated as potential isotopologues from the forty-fivev reference compounds in question. Overall, 212 TP cases, one FN case, and one FP case were found for the classification model, Resulting in an average TP_r of 99.8% and an FD_r of 0.5%. The FN case was caused by an 0.011 Da mass error between the monoisotopic and isotopologue mass, which is larger than the minimum mass error (i.e., 0.01 Da) assumed for the true negative cases that are used for training the model. As for the FP case, the detected isotopologue mass was 155.068 m/z and the monoisotopic parent ion mass was 152.072 m/z. If the decreasing intensity for less likely isotopologues would have been taken into account, this ion would not have been included due to the absence of the isotopologues with a higher probability (e.g., 153.068 and 154,075 m/z, Figure S7). From this, it can be concluded that the classification model can also be used for real data.

For the mass difference method, a total of 203 TP, 10 FN, and 13 FP cases were found,

corresponding to an average TP_r and FD_r of 96.3 and 4.8%. For these cases, all FNs were caused by a mass error larger than 0.01 Da and all FPs were caused by the same reason as the FP of the classification model. Across multiple datasets a signal at 304.182 m/z was identified as an isotopologue of codeine, for which the monoisotopic mass was 300.159 m/z. Only in some cases, an isotopologue at 301.163 m/z was detected, which would still mean that there were no isotopologues with an isotopologue mass depth of 2 or 3 present with higher intensities than the signal at 304.182 m/z. To conclude, the classification model had a higher TP_r and lower FD_r than the mass difference method. However, if the decreasing intensity with lower isotopologue probabilities would have been taken into account, the methods would both have had an FD_r of 0.0%.

Potentials and Limitations

The classification model provides a good alternative approach for the detection of isotopologues, requiring no information on the molecular formula or arbitrary thresholds. However, it should be noted that the classification model is unable to distinguish between isotopologues coming from different chemicals or signals with the same monoisotpic mass. This would require prior separation such as chromatography. Besides the reduction of total number of features for identification, correct isotopologue identification can also assist in accurate molecular formula assignment. When multiple formula's are possible for a monoisotopic mass, the isotopic patterns can be predicted and compared with the detected isotopologues masses to eliminate less likely candidates. Lastly, the model was built based on isotopic distributions with a tree depth of six, meaning that it might not be able to correctly classify ions with more than 6 isotopologues if these ions would be detected at all due to their low occurrence probabilities. However, if required, the EMDforIso package enables the user to retrain the classification model using different training sets and parameters.

Conclusion

This manuscript demonstrated the potential of using elemental ratios for the detection of isotopologues. The classification model that was constructed based on the elemental ratios CO, CN, CCl, CS, CF, and CH, showed good performance for both theoretical isotopic patterns as well as real wastewater influent samples. For the theoretical mono-iso pairs, when assuming no error, the classification model outperformed the mass difference method with a TP_r of 99.0% and FP_r of 1.8% compared to a TP_r of 16.2% and an FP_r of 0.02%. As for the wastewater influent samples, the classification model, with a TP_r of 99.8% and FD_r of 0.5%, performed better than the mass difference method, with a TP_r of 96.3% and FD_r of 4.8%. However, if a decreasing intensity for a lower probability isotopologue was taken into account, both methods would have had an FD_r of 0.0 %.

Acknowledgement

The authors thank the wastewater treatment plants that assisted in the collection of the wastewater influent samples, the Chemometrics and Advances Separation Team (CAST) for their insights, and the authors gratefully acknowledge the financial support from the Australian Research Council ARC Discovery Project (DP190102476) and Linkage Project(LP150100364). Finally, the Queensland Alliance for Environmental Health Sciences, The University of Queensland, gratefully acknowledges the financial support of the Queensland Department of Health

Supporting Information Available

Information on the presence of the elemental ratios for the chemicals in the DDS-Tox database, an overview of correlation coefficients for the different elemental ratios between the EMD_{mono} and EMD_{iso} values with scatter plots for the two most extreme correlations,

receiver operator curves for the classification model and mass difference method used for the selection of the $score_{EMD}$, a reference compound list used for the performance assessment of the classification model and mass difference method on wastewater influent samples, and an example of FP detected isotopologue for the classification model.

Author Information

Corresponding Author:

Saer Samanipour

Van 't hoff institute for molecular sciences (HIMS),

University of Amsterdam,

the Netherlands

Email: s.samanipour@uva.nl

Denice van Herwerden

Van 't hoff institute for molecular sciences (HIMS),

University of Amsterdam,

the Netherlands

Email: d.vanherwerden@uva.nl

References

(1) Bastian, S.; Youngjoon, J.; Sarit, K.; Amy, H. L.; Pradeep, D.; Jake, O.; Maria Jose, G. R.; Sara, G. G.; Jochen, M. F.; Kevin, T. V.; Saer, S. An assessment of Quality Assurance/Quality Control Efforts in High Resolution Mass Spectrometry Non-Target Workflows for Analysis of Environmental Samples. *TrAC Trends in Analytical Chemistry* **2020**, *133*, 116063.

(2) 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.* **2015**, *407*, 6237–6255.

(3) Werner, E.; Heilier, J.-F.; Ducruix, C.; Ezan, E.; Junot, C.; Tabet, J.-C. Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends. *J. Chromatogr. B* **2008**, *871*, 143 – 163, Hyphenated Techniques for Global Metabolite Profiling.

(4) Samanipour, S.; Reid, M. J.; Bæk, K.; Thomas, K. V. Combining a Deconvolution and a Universal Library Search Algorithm for the Nontarget Analysis of Data-Independent Acquisition Mode Liquid Chromatography-High-Resolution Mass Spectrometry Results. *Environ. Sci. Technol.* **2018**, *52*, 4694–4701.

(5) Samanipour, S.; Kaserzon, S.; Vijayasarathy, S.; Jiang, H.; Choi, P.; Reid, M. J.; Mueller, J. F.; Thomas, K. V. Machine learning combined with non-targeted LC-HRMS analysis for a risk warning system of chemical hazards in drinking water: A proof of concept. *Talanta* **2019**, *195*, 426 – 432.

(6) Brack, W.; Hollender, J.; de Alda, M. L.; Müller, C.; Schulze, T.; Schymanski, E.; Slobodnik, J.; Krauss, M. High-resolution mass spectrometry to complement monitoring and track emerging chemicals and pollution trends in European water resources. *Environ. Sci. Eur.* **2019**, *31*, 62.

(7) Stravs, M. A.; Hollender, J. Open-source workflow for smart biotransformation product elucidation using LC-HRMS data. *Anal. Chem.* **2017**, *86*, 6812–6817.

(8) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2012**, *84*, 283–289.

(9) Röst, H. L. et al. OpenMS : a flexible open-source software platform for mass spectrometry data analysis. *Nat. methods* **2016**, *13*, 741–748.

(10) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinf.* **2010**, *11*.

(11) Ort, C.; Lawrence, M. G.; Rieckermann, J.; Joss, A. Sampling for pharmaceuticals and personal care products (PPCPs) and illicit drugs in wastewater systems: Are your conclusions valid? A critical review. *Environ. Sci. Technol.* **2010**, *44*, 6024–6035.

(12) Samanipour, S.; O'Brien, J. W.; Reid, M. J.; Thomas, K. V. Self adjusting algorithm for the nontargeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data. *Anal. Chem.* **2019**, *91*, 10800–10807.

(13) Chambers, M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918–920.

(14) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology* **2019**, *12*, 100096.

(15) Łacki, M. K.; Startek, M.; Valkenborg, D.; Gambin, A. IsoSpec: Hyperfast Fine Structure Calculator. *Anal. Chem.* **2017**, *89*, 3272–3277.

(16) Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. Accelerated isotope fine structure calculation using pruned transition trees. *Analytical Chemistry* **2015**, *87*, 5738–5744.

(17) Alygizakis, N. A.; Samanipour, S.; Hollender, J.; Ibáñez, M.; Kaserzon, S.; Kokkali, V.; Van Leerdam, J. A.; Mueller, J. F.; Pijnappels, M.; Reid, M. J.; Schymanski, E. L.; Slobodnik, J.; Thomaidis, N. S.; Thomas, K. V. Exploring the Potential of a Global Emerging Contaminant Early Warning Network through the Use of Retrospective Suspect Screening with High-Resolution Mass Spectrometry. *Environ. Sci. Technol.* **2018**, *52*, 5135–5144.

(18) van Herwerden, D.; Samanipour, S. Dataset for: probabilistic classification model for isotope detection in high-resolution mass spectrometry. 2021; `https://figshare.com/articles/dataset/DDS-Tox_isotope_distributions/16559517/2`.

(19) Albon, C. In *Machine Learning with Python Cookbook*; Roumeliotis, R., Bleiel, J., Eds.; O'Reilly Media, Inc., 2018.

(20) Brereton, R. G. *Applied Chemometrics for Scientists*; John Wiley & Sons, 2007.

Graphical TOC Entry