



LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Towards Machine Learning-Driven Mass Spectrometric Identification of Trichothecenes in the Absence of Standard Reference Materials

B. P. Mayer, M. L. Dreyer, M. C. Prieto Conaway,
C. A. Valdez, T. Corzett, R. Leif, A. M. Williams

December 12, 2023

Analytical Chemistry

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Towards Machine Learning-Driven Mass Spectrometric Identification of Trichothecenes in the Absence of Standard Reference Materials

Brian P. Mayer*, Mark L. Dreyer, Maria C. Prieto Conaway, Carlos A. Valdez, Todd Corzett, Roald Leif, Audrey M. Williams

Forensic Science Center, Lawrence Livermore National Laboratory, 7000 East Avenue L-090, Livermore, California, 94550.

ABSTRACT: While a significant body of work exists on the detection of commonly known trichothecene toxins, biological, environmental, and other transformational processes can generate many under-characterized and unknown modified trichothecenes. Lacking both analytical reference standards and associated mass spectral databases, identification of these modified compounds reflects both a challenge and a critical gap from forensic and public health perspectives. We report here the application of machine learning (ML) techniques towards identification of discriminative fragment ions from mass spectrometric data that can be exploited to detect evidence of type A and B trichothecenes. The goal of this work is to establish a new, complimentary method for the identification of unknown, though structurally similar trichothecenes by leveraging objective ML techniques. Discriminative fragments derived from a series of gradient boosted machine learners are then used to develop ML-driven precursor ion scan (PIS) methods on a triple quadrupole mass spectrometer (QQQ) for screening of “unknown unknown” trichothecenes. Specifically, we apply the PIS method to a laboratory-synthesized trichothecene, a first step in demonstrating the power of alternative methods for screening that rely on machine learning-driven mass spectrometric methods.

Trichothecenes are secondary metabolite toxins produced mainly by the *Fusarium* genus of fungi and are frequently found as contaminants in the food supply chain, affecting approximately 25% of the global food and feed crop output.^{1,2,3} This broad class of compounds includes approximately 180 known analogs, many of which are toxic to both humans and animals.^{4,5,6}

These toxins are sesquiterpenes, containing a common tricyclic 12,13-epoxytrichothec-9-ene moiety. They are divided into four classes according to structure: type A has an -OH, ester, or hydrogen at the C-8 position, type B has a C-8 keto group, type C has a second epoxide group, and type D includes more complex, macrocyclic compounds. Types A and B are produced by *Fusarium* species, while type C and D are not.⁷ The figure in Table 1 shows the generalized structure of type A and B trichothecenes, while the table itself highlights the structural differences among the studied toxins.

Though food and feed regulations restrict exposure to several key trichothecenes, potential exposure to toxins that have been chemically modified by plant or animal metabolic processes remains an acute concern. Major metabolic pathways such as hydrolysis, hydroxylation, deepoxidation, and conjugation⁸ have the potential to produce a large number of compounds from a small number of secondary metabolites; and while many of these compounds have been characterized, it is reasonable to assume many more so-called “modified trichothecenes” exist that are under-characterized or altogether unknown.

Conventional analytical screening techniques routinely miss and underestimate the total toxin content in contaminated food due to a lack of metabolic knowledge and poor availability of analytical standards of modified trichothecenes. Developing new and alternative analytical chemistry workflows that screen for trichothecene presence is therefore

urgently needed for forensic analysis of biological and agricultural samples.

Comparison of mass spectral data of unknowns to those in databases is the most common approach for correlating structural relatives of known compounds. This method is only applicable to “known unknowns” for which database data of the related compounds are available. Several alternative but complementary strategies have been developed to identify “unknown unknown” toxins using untargeted approaches. The retrospective analysis of high-resolution mass spectrometry (HRMS) data has been used most frequently.^{9,10} Another nontargeted strategy has involved data independent MS/MS analysis for screening and quantitation of secondary metabolites in green tea¹¹ and cruciferous vegetables.¹²

A final example is the use of precursor ion scans on triple quadrupole mass spectrometers (QQQ). Fragment ions common to compounds sharing structural features are scanned for in Q3 of the mass spectrometer, then Q1 is interrogated for the precursor ions producing those fragments. By screening for fragment ions claimed to be specific to type A trichothecenes, several toxin analogs not previously described from an *F. sporotrichioides* strain were reported.⁷

The work cited above is to our knowledge the only publication focusing on the direct determination of MS features common to a subset of trichothecenes. In this and related literature, however, features deemed discriminative are generally determined subjectively. These diagnostic fragment ions are frequently those most easily discerned by eye from retrospective data analysis. Meanwhile, low intensity ions that are potentially statistically relevant are not discussed, likely due to the difficulty of manually extracting meaningful, low intensity features from complex datasets.

Broadly speaking, methods like multivariate statistical analysis and other techniques common to machine learning (ML) make possible the objective analysis of complex, large

data sets. They can be used to identify predictors (i.e., measured variables) statistically relevant for discrimination and allow for classification of unknown compounds based on training data from known chemicals. Much of this work, however, has largely been directed towards metabolomics and related fields.^{13,14,15}

The work reported herein focusses on supervised statistical analysis of LC-MS/MS data from a panel of commercially available type A and type B trichothecenes. Mass spectral features differentiating the two toxin classes are identified from ML learners trained on high-resolution accurate mass data and are used to direct a new detection strategy for modified trichothecenes. In the current research's context, the goal of applying machine learning is not to generate and apply a classifier to unknown data as is often done. Nor is it the intent to provide a more sensitive detection method as compared to orbitrap and other high-resolution MS techniques. Instead, it is to identify fragments important for trichothecene classification and develop a broadly implementable MS method leveraging those discriminatory predictors, which can more specifically target unknown trichothecenes and compliment other MS-based techniques. Limitations of the current strategy and opportunities for future study are also discussed.

EXPERIMENTAL SECTION

Chemicals. T-2, HT-2, DAS, 3-AcDON, DON, 15-AcDON, 15-AcS, NEO, T-2-4OH, T-2-3OH, and NIV were received as powders from Cayman Chemical (Ann Arbor, MI, USA). FX was also received from Cayman but as a 1 mg/mL solution in dichloromethane. Additional samples of T-2, HT-2, and DAS were purchased from Toronto Research Chemicals as powdered material. DEDON and DON-3G were obtained from Sigma-Aldrich (St. Louis, MO, USA) as 50 µg/mL solutions in acetonitrile. When possible, toxins were prepared as 1 mg/mL stock solutions in LC-MS-grade acetonitrile (Optima™ grade, Fisher Scientific). Tri-ACDON was synthesized in house, and details can be found in Supporting Information.

Sample Preparation. For LC-HRMS analysis, stock solutions were diluted with 50% (v/v) acetonitrile/Milli-Q water to a concentration of 100 µg/mL, then further diluted with 50% (v/v) acetonitrile/Milli-Q water to 10 µg/mL working stocks. Working solutions were diluted to 500 ng/mL in 2:1:1 water:acetonitrile:methanol for analysis.

HRMS Analysis. MS acquisition was performed on a Thermo Scientific Q Exactive HF-X mass spectrometer operated using heated electrospray ionization in positive mode. The MS was used with high resolution accurate mass to ≤ 3 ppm. A 3 min MS experiment was composed of a full MS spectrum (m/z 100–1500) at a resolving power setting of 240,000 (full width at half maximum (FWHM) at m/z 200) followed by MS/MS acquisition (minimum m/z 50) at resolving power of 45,000 (FWHM at m/z 200). Final collision energies for data used for subsequent machine learning were chosen so 1% to 10% of the parent ion remained. In most cases, the toxins' $[M+NH_4]^+$ adduct provided the best signal-to-noise for the direct infusion runs. $[M+H]^+$ data of 3-AcDON, DON, DEDON, and NIV were used, however, which was deemed acceptable since MS/MS fragmentation of ammonium and hydrogen adducts were observed to be highly similar. Data were exported to a .csv using Thermo's Xcaliber Qual Browser software by averaging over 20 consecutive acquisition time points and using the 'Export to Excel' function. No other processing of the data was performed apart from seven-point Gaussian smoothing prior to export. Data for statistical analysis were generated via infusion in triplicate, resulting in 42 MS spectra. A table of collision energies and adducts is given as Table S11 in Supporting Information.

Precursor Ion Scans. Precursor ion scan experiments were conducted using a Thermo Scientific Altis triple quadrupole MS. In these experiments, specific fragments are scanned in the Q3 of a triple quadrupole, and Q1 is interrogated for the precursor ion connected to those fragments. Mobile phases used were A = water/5 mM ammonium formate/0.1% formic acid and B = methanol/5 mM ammonium formate/0.1% formic acid. The solvent gradient was 12 min long using a one-minute initial hold at 0% B, a 0% to 50% B linear gradient over 2.25 minutes, a 50% to 65% linear gradient over 2.75 min, a 65% to 100% linear gradient over 1 min, a 2 min hold at 100% B, followed by reequilibration for 3 min at 0% B. 3 µL of the 500 ng/mL samples were injected into the LC-MS/MS for analysis.

Multivariate Statistical Analysis. The machine learning methods used in this work were adapted from those applied to several studies we have previously published.^{16,17,18} HRMS data were imported and processed by machine learning code using the R programming language.¹⁹ LC-HRMS data were binned into 0.05 amu bins. Each mass spectrum was then normalized to its base (i.e., most intense) peak (base peak normalized to intensity = 100). Packages used in the analyses were caret,²⁰ corrplot,²¹ MASS,²² doMC,²³ Boruta,²⁴ reshape, gplots, RColorBrewer, and dplyr. Data preprocessing, reduction and filtering, and generation of data plots were performed using algorithms native to R and the above packages.

Variable reduction was performed using the Boruta algorithm. This supervised feature selection method removes irrelevant predictors. The remaining predictors were reordered in descending variable importance. Classification models were developed using the gradient boosted trees (GBM) algorithm.²⁵ We continue to use this method, as it – along with other feature selection algorithms – is generally recognized to be robust against overfitting, which plagues limited data sets.²⁶ See further discussion below.

RESULTS AND DISCUSSION

All Relevant Feature Selection. Preprocessing of the LC-HRMS data was performed as described in the Methods section above. Data were binned to small m/z ranges to align the spectra. Boruta analysis, a random forest-based feature selection algorithm, was then applied to identify those predictors not relevant for classification of trichothecenes. This process serves to reduce the dimensionality of the data by discarding non-discriminative predictors while retaining all potentially relevant predictors. This and related procedures are commonly used when analyzing “high dimension, low sample size” (HDLSS) datasets common in chemistry. Note the algorithm has no adjustable parameters..

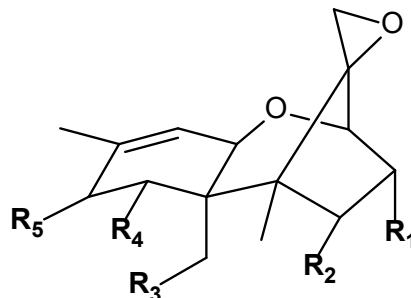
Results from training the Boruta algorithm (at a p -value of 0.01) to discriminate between type A and type B trichothecenes showed a 99.3% reduction in the set of 9187 predictors (i.e., MS/MS data bins), resulting in only 67 relevant predictors. This removal of statistically irrelevant predictors reduces the potential of model overfit, avoids introducing unnecessary noise, and allows for more tractable interpretation of results..

Classification Model Optimization. After conducting Boruta dimensional reduction, a gradient boosted tree classification model was optimized on the remaining 67 predictors to discriminate between type A and type B trichothecenes. The tunable GBM parameters of minimum terminal node size and interaction depth were observed to have little effect on model performance and were set to one (see Figure S1 in Supporting Information). Shrinkage, however, impacted cross validated accuracy and was set at $s = 0.005$ as a compromise between good generalization and having too

slow a learning rate. With these parameters optimized, the optimum number of boosting iterations was determined to be 24. This number comes from ~1.5 times the number of trees at which the standard deviation of CV accuracy went

to zero at $s = 0.005$. Standard deviation was calculated using results of 100 individual models for a given number of trees.

Table 1. (Top) Generalized structure for type A and type B trichothecenes. Moieties for R1 through R5 and their relation to specific toxins are given in the table below. R5 represents a moiety at the C-8 position of the base trichothecene structure. (Bottom table) Summary of the trichothecene panel investigated in this study. Refer to the generalized toxin above the table to derive the structure of a specific trichothecene. Note type B trichothecenes all have a ketone group at the R5 position. "OAc" and "Oisoval" refer to acetyl and isovaleryl (3-methylbutanoate) groups, respectively.



Trichothecene	Type	R ₁	R ₂	R ₃	R ₄	R ₅
T-2 toxin (T-2)	A	OH	OAc	OAc	H	Oisoval
HT-2 toxin (HT-2)	A	OH	OH	OAc	H	Oisoval
Diacetoxyscirpenol (DAS)	A	OH	OAc	OAc	H	H
T-2 tetraol (T-2-4OH)	A	OH	OH	OH	H	OH
T-2 triol (T-2-3OH)	A	OH	OH	OH	H	Oisoval
Neosolaniol (NEO)	A	OH	OAc	OAc	H	OH
15-Acetoxyeirpenol (15-AcS)	A	OH	OH	OAc	H	H
Deoxynivalenol (DON)	B	OH	H	OH	OH	=O
3-Acetyldeoxynivalenol (3-AcDON)	B	OAc	H	OH	OH	=O
15-Acetyldeoxynivalenol (15-AcDON)	B	OH	H	OAc	OH	=O
Fusarenon-X (FX)	B	OH	OAc	OH	OH	=O
Nivalenol (NIV)	B	OH	OH	OH	OH	=O
Deepoxydeoxynivalenol * (DEDON)	B	OH	H	OH	OH	=O
DON 3-glucoside ** (DON-3G)	B	OGlu	H	OH	OH	=O

* A deoxynivalenol (DON) metabolite where the epoxy moiety has been cleaved to give an ethene group; ** A DON metabolite where a glucoside group ("OGlu") has condensed with the hydroxyl moiety at the R₁ position.

A type A/B classification model was trained using these optimized values. Cross-validated accuracies of the training set showed these training data could be classified accurately for all MS spectra/samples considered (accuracy and kappa both equaled 1.0). We refer to this model as the "optimized model."

Gradient boosted trees are attractive algorithms because they offer a natural way of quantifying the relative influence of each feature on the final model. This influence is most frequently parameterized through "variable importance" (VI). The variable importance of predictors revealed this optimized type A/B model only required a single fragment ion to discriminate between the two toxin types, $m/z_{bin} 165.1$. That is, the VI of $m/z_{bin} 165.1$ was shown to be 100; no other predictors were needed to differentiate between type A and type B toxins.

Discriminative Fragment Ions. Upon examination of the raw MS data, the corresponding fragment ion accurate mass was found to be $m/z 165.0908 (\pm 1.5 \text{ ppm})$. No other significant exact masses fell in this 0.05 m/z range in any of the HRMS data. Considering probable formulae for this exact mass, this fragment most likely corresponds to $[C_{10}H_{12}O_2 + H]^+$. Considering shared structural features of type B trichothecenes, this fragment likely includes the R₄ and R₅ groups and the cyclohexene ring of the toxins.

Examination of the normalized binned data (see Figure 1) for $m/z 165.1$ reveals the predictor strongly indicates a type B trichothecene, with only trace normalized signals (< 1.0) for type A toxins. This plot also suggests why the classification model only requires a single predictor to discriminate between the two toxin types.

The plotted m/z_{bin} 165.1 data in Figure 1 also provide the first suggestion that manual curation of data is unlikely to identify statistically significant discriminative fragment ions from MS/MS data. Firstly, the nonzero normalized intensity for type A toxins may cause m/z_{bin} 165.1 to be prematurely ruled out through manual data analysis. Secondly, the normalized responses of m/z_{bin} 165.1 for type B toxins span a significant range of values: from 0.8 for NIV to 20.5 for DEDON. This variability may also lead the analyst to exclude m/z_{bin} 165.1 as important. Figure 1, however, visualizes the potential predictive power of this fragment, which the classification model objectively found highly statistically significant.

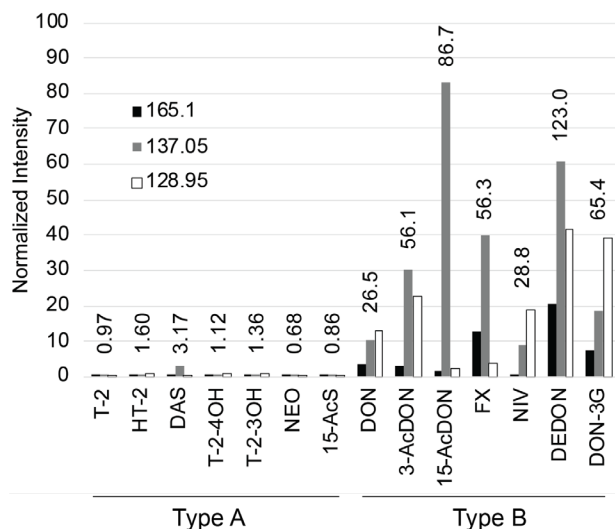


Figure 1. Normalized intensities of the top three ions (m/z_{bin} 165.1, 137.05, and 128.95) discriminative for type B toxins. Fragment m/z_{bin} 165.1, the most important predictor, was the only fragment needed by the optimized model to differentiate type A from type B toxins. Fragments m/z_{bin} 137.05 and 128.95 were the derived from the modified model (see further discussion in the text). Values are normalized intensity averages over triplicate spectra; standard deviation of measurement not shown to reduce visual clutter. Sums of the three ions' responses are given vertically above the bars.

Still, additional ions would be useful to help minimize false positives in identification/classification, akin to quantification by MS techniques, which relies on “qualifier” ions in addition to the principal “quantifier” ion. We envisioned an approach to identify such “ML qualifying” ions by modifying the optimized classification model by increasing greatly the number of trees, in a sense forcing an overfit. This intentionally non-optimal learner identified additional fragments – again quantified through variable importance – potentially diagnostic for classification. 1500 trees were chosen for this overfitted model, and we refer to this model as the “modified model.”

Normalized responses for the top three predictors (assessed through their variable importance in the modified model, VI) are given in Figure 1. These discriminative fragment ions, in addition to m/z_{bin} 165.1 (VI = 83.2, the optimized model also determined this to be the most important variable), were 137.05 (VI = 7.0) and 128.95 (VI = 2.20). All three m/z_{bin} were strongly indicative of type B

trichothecenes and shows m/z_{bin} 137.05 and 128.95 are both good discriminative “ML qualifier” ions.

All the ions considered so far have been exclusively indicative of type B trichothecenes. It seems understandable, then, that only a single fragment be needed for a classifier aimed at two classes. For example, if the relative intensity of a fragment associated with type B toxins is low, then it might be assumed the toxin is of type A. This logic does not, however, extend easily to analyses aimed at screening. Real, complex samples will contain many compounds at various concentrations, and methods for identifying the presence of a toxin will require screening multiple fragment ions. Ideally, a useful screening method will also explicitly incorporate type A-specific fragments as well.

Returning to the modified model's VI metrics, the fourth most important variable was found to correlate with type A toxins. Normalized responses for m/z_{bin} 217.15 (VI = 1.55) are shown in Figure 2. Note while the relative responses are relatively low (< 2.5%), this fragment ion displays good discriminative ability. There exists, however, significant signal for the type B DEDON. HRMS data revealed the accurate masses of the fragments were the same for all compounds: m/z 217.1124 with no other competing peaks in this bin. This mass corresponds to $[C_{14}H_{16}O_2 + H]^+$ at 217.1123 (mass error = 0.2 ppm). Though the accurate mass is the same for DEDON and type A toxins, screening for other type B-specific fragments would help bolster a putative trichothecene identification and toxin class assignment.

Using the modified model, three additional, type A “qualifier” fragment ions were identified, m/z_{bin} 105.05, 169.1, and 215.1. Responses are shown in Figure 3. All three “qualifier” ions are more intense than the “best” predictor, m/z_{bin} 217.15, from Figure 2; but their discriminative importance is not as large, particularly considering the poor overall response from 15-AcS and the positive, often intense responses for type B toxins.

Firstly, fragment m/z_{bin} 215.1 was found to be diagnostic for most type A toxins and was often the base peak of the MS data. Notable exceptions were the two scripenol-based toxins, DAS and 15-AcS. The work of González-Jartín⁷ also demonstrated this fragment is indicative of type A trichothecenes, but the present data suggest scripenol-based toxins would go undetected if only m/z 215.1 was used as a screening fragment. The same is true for the two other diagnostic ions González-Jartín, et al. suggested: m/z 245.1172 and 197.0961. All three ions are not present in DAS and 15-AcS data. This fact is likely related to the hydrogen atom at the R₅ position of these two compounds. The other five compounds all have an OH in that position either natively (NEO and T2-4-OH) or from losing an isovaleryl group during MS/MS fragmentation (T-2, HT-2, and T2-3-OH). These differences are expected to significantly alter the fragmentation chemistry of DAS and 15-AcS relative to the other five type A trichothecenes, as a R₅ hydroxyl can be liberated through a loss of water, which is not possible for either DAS or 15-AcS. The dissimilar fragmentation of these two compounds emphasizes the

importance of both identifying and employing a panel of diagnostic fragments using larger toxin panels.

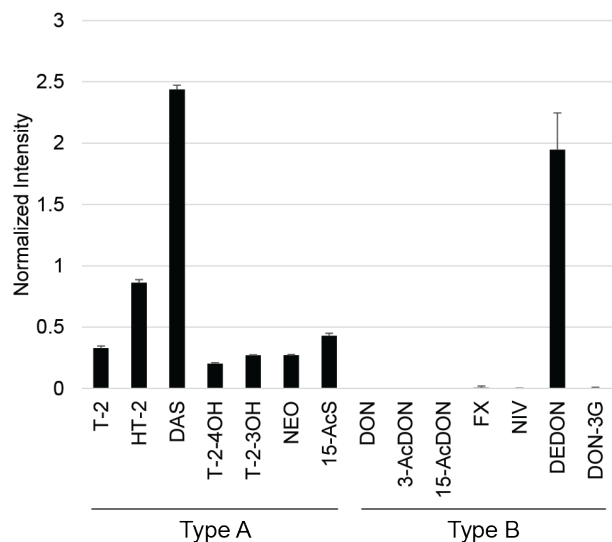


Figure 2. Normalized responses of predictive fragment ion $m/z_{\text{bin}} 217.15$, which correlates well with type A trichothecenes. DEDON is an exception, but it can be confirmed using type B-specific fragment ions. Intensities are averages of the triplicate MS data.

Precursor Ion Scan-based screening of Trichothecenes. Using the diagnostic fragment ions identified above, mixtures of known trichothecenes were investigated using precursor ion scans (PIS) on a QQQ-MS. The goal of these experiments was to investigate the extent to which the ML results can be used to drive the creation of practical, implementable MS-based screening techniques. A QQQ was employed for the ability of its PIS experiment to provide parent masses from targeted product ions when parent compounds are not *a priori* known.

PIS methods were generated for each of the seven diagnostic fragments derived from the modified classification model summarized in Table 2. The table gives the exact masses for the fragments (from the HRMS data) in addition to the most probable empirical formulae. For discussion that follows, also refer to the PIS chromatograms shown in Figures SI2 and SI3 of Supporting Information. Results from applying these methods to separate type A and type B mixtures suggest the methods display good potential for screening for trichothecenes in general. For the type A method, the top three ions (the preferred m/z 217.1 and two qualifiers, m/z 105.1 and 169.1) show signals for all seven type A toxins. Only DAS and 15-AcS lacked visible signal for m/z 215.1, consistent with the HRMS data shown in Figure 3.

In general, detectable signal, though low compared to those from type A toxins, was observed for some type B toxins using the type A method. For example, there is a significant peak at 4.20 min in the m/z 217.1 PIS data (Figure SI2.B1). This feature belongs to FX. Note from Table 2 the targeted high resolution fragment for this screen is m/z 217.1224, or $[\text{C}_{14}\text{H}_{16}\text{O}_2 + \text{H}]^+$. Cross-referencing the HRMS data shows response for FX at m/z 217.0865,

however, corresponding to $[\text{C}_{13}\text{H}_{12}\text{O}_3 + \text{H}]^+$ (mass error = 0.5). The use of a nominal mass instrument would therefore result in a false positive type A identification if only $m/z_{\text{bin}} 217.1$ were targeted.

Table 2. Trichothecene type-specific fragment ion exact masses determined from raw HRMS data and their corresponding molecular formulae, theoretical m/z , and mass error.

Type	Meas. m/z	Fragment ion empirical formula	Theor. m/z	Mass error (ppm)
A	217.1224*	$[\text{C}_{14}\text{H}_{16}\text{O}_2 + \text{H}]^+$	217.1223	0.5
	105.0700	$[\text{C}_8\text{H}_8 + \text{H}]^+$	105.0699	1.0
	169.1010	$[\text{C}_{13}\text{H}_{12} + \text{H}]^+$	169.1012	-1.2
	215.1066	$[\text{C}_{14}\text{H}_{14}\text{O}_2 + \text{H}]^+$	215.1067	-0.5
B	165.0911*	$[\text{C}_{10}\text{H}_{12}\text{O}_2 + \text{H}]^+$	165.0910	0.6
	137.0598	$[\text{C}_8\text{H}_8\text{O}_2 + \text{H}]^+$	137.0597	0.7
	128.9536	--**	--	--

* “Preferred” screening fragment ion; other ions are referred to as “ML qualifier” ions. ** No empirical formula assignment could be made using carbon, hydrogen, oxygen, nitrogen, sodium, and potassium.

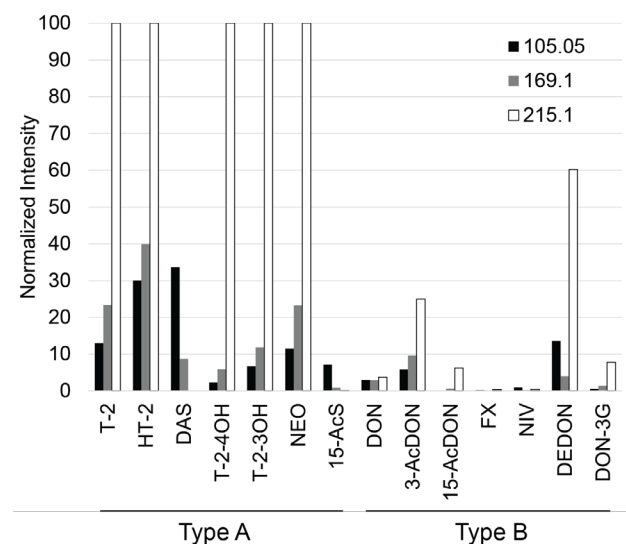


Figure 3. Normalized binned intensities for qualifying fragment ions indicative of type A trichothecenes ($m/z_{\text{bin}} 105.05$, 169.1, and 215.1) from the modified classification model. Values are averages over the triplicate spectra; standard deviation of measurement not shown to reduce clutter.

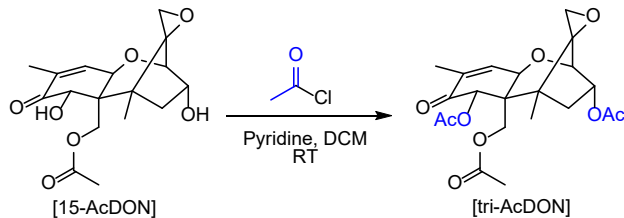
This cross-class issue worsens considering the inherently excellent sensitivity of QQQ instrumentation, increasing the likelihood of false positive identifications. As a prime example, HRMS data for $m/z_{\text{bin}} 105.1$ show all 14 toxins studied generate the fragment. These data also show the fragment has the same exact mass across all trichothecenes: m/z 105.0700 ($[\text{C}_8\text{H}_8 + \text{H}]^+$). This is not the only time this issue is observed. The HRMS data, for example, also show positive, cross-class detection of m/z 215.1066 for many type B trichothecenes.

The rate of false positive identification appears to worsen when considering the type B PIS method (see Figure S13). First, while detection of all four type B toxins was positive (DEDON, NIV, and DON-3G were not present in the test mix), responses like those of DON (3.73-3.74 min) are quite low. This issue is compounded by the fact the use of the nominal mass QQQ negates the diagnostic power of m/z 128.9536, detecting responses for all 11 toxins studied using the PIS methods.

A Global Approach to Unknown Trichothecene Screening. All seven diagnostic fragments were used in combination to develop a master trichothecene screening panel. This capability is envisioned as a first step of a workflow seeking to identify trichothecene toxins with subsequent analyses being performed as necessary to further elucidate unknown trichothecene structures (e.g., full-scan and/or targeted HRMS acquisition).

Here we demonstrate this idea on a laboratory-synthesized trichothecene, 3,7,15-triacetyldeoxynivalenol (tri-AcDON), which serves to mimic an undiscovered or emergent type B toxin (Scheme 1 and further information in Supporting Information). A crude, unpurified reaction mixture of tri-AcDON (5 $\mu\text{g/mL}$ of crude mixture in LC-grade water) was analyzed using the seven-fragment PIS screening method on a LC-QQQ (see details on the method in the Methods section above). LC-QQQ chromatograms acquired using the PIS screen are given in Figure 4.

Scheme 1. General synthetic approach for generating tri-AcDON. Synthesis details are given in Supporting Information.



The total ion chromatogram of the tri-AcDON reaction mixture (Figure 4a) shows a complex reaction mixture with many chromatographic features. An extracted ion chromatogram (XIC) of the expected $[M+H]^+$ of m/z 423.1650 shows the product is clearly visible at 5.55 min (Figure 4b), demonstrating successful synthesis of the fully acetylated type B trichothecene, tri-AcDON.

PIS data using the seven diagnostic fragments follow in Figure 4c-i. These chromatograms show strong, interference free signals at the 5.55 min retention time associated with tri-AcDON. Fragments m/z_{bin} 129.1 and 165.1 show weaker response, but this behavior is consistent with that of its parent compound 15-AcDON, which displayed similar weak intensities compared to the third type B fragment, m/z_{bin} 137.1 (see Figure 1).

Note in the PIS chromatograms in Figure 4c-i a second peak at a retention time of 5.28 min. This peak clearly appears for five of the seven PIS scans, suggesting strongly this compound is a second trichothecene related to 15-AcDON. The most likely candidate was a diacetylated relative ($[M+H]^+ = m/z$ 381.1544) since the reaction mixture

was analyzed as a crude material (i.e., unpurified). An XIC for the parent ion (Figure 4j) shows a clear peak at 5.28 min, consistent with a diacetylated trichothecene analogue. A smaller peak appears at 5.55 min, which was determined to be an in-source fragment of tri-AcDON.

The success of identifying clear PIS responses for both the intended tri-AcDON and the reaction diacetylated reaction byproduct, demonstrates the potential power of a ML-driven mass spectrometric screening method. Particularly notable is the performance of the method applying a low-resolution QQQ instrument to a crude reaction mixture containing byproducts, excess reagents, and other confounding impurities. At first blush, it may appear problematic that clear PIS peaks were observed for type A-specific fragments, despite tri-AcDON and its byproduct being both type B toxins. The data in Figure 4, however, show this initial study has provided is a method that can help identify novel trichothecenes warranting further study, for example, by transferring LC-QQQ PIS retention time to a time-of-flight or orbitrap instrument targeting the parent ion mass from the PIS data. The power of the method can only be increased by using a high-resolution instrument, which filters out MS interferants by targeting specific masses and leveraging diagnostic fragment mass defects.

CONCLUSIONS

This work details the application of machine learning methods towards the creation of ML-driven mass spectrometric methods that can be applied towards identifying spectrometric evidence of “unknown unknown” trichothecene mycotoxins.

Diagnostic fragments were discovered that could statistically differentiate between type A and type B trichothecenes. For an optimized classifier, it was found that only a single fragment (m/z 165.0908, $[C_{10}H_{12}O_2 + H]^+$) was needed for classification. To identify additional ions that could be used as “ML qualifying” ions, overfitted, “modified” models were employed. An additional six ions were identified from these efforts that were then used to develop a PIS screening method on a triple quadrupole mass spectrometer. Using an unpurified, laboratory-synthesized trichothecene, tri-AcDON, it was shown that the PIS screen could be used to reveal evidence of trichothecenes, including an unexpected, diacetylated reaction byproduct.

The data collected for this study emphasize the benefits of statistical analysis of large complex data sets, as the response of the most important predictive ions were found to be relatively low response (e.g., m/z_{bin} 165.1 vis-à-vis m/z_{bin} 137.05 for type B, or m/z_{bin} 217.15 vis-à-vis m/z_{bin} 215.1 for type A). This fact would further suggest objective data analysis should be preferred over manual data curation, as important fragments would almost certainly be missed with the latter.

The extent to which the developed ML-driven fragment screen is more broadly applicable depends on the specifics of the compounds used to train and extract fragments from the ML model. We do believe, however, establishing the model using 14 diverse trichothecenes makes it one of the most objective, applicable screening tools yet developed

for this toxin class. Developing a model based on larger classes of trichothecenes is expected to only further aid in classification reliability and accuracy.

Future work seeks to expand the robustness and forensic practicality of this model by firstly developing a learner aware of out-of-class MS data. This could involve including mass spectral data from other sources such as 1) other toxin classes, 2) other plant metabolites, 3) or randomly chosen MS data sampled from high quality HRMS libraries. This would certainly change the model itself but would not invalidate the screening capability developed herein. If the model used here were actually used to automatically identify and classify MS data of unknown unknown trichothecenes, then a refined model trained on a large library of non-trichothecene spectra would be critical to increasing model performance, applicability, and accuracy.

The robustness of the current models can be further enhanced by acquiring HRMS training data over a range of collision energies thereby creating a classification model independent of this experimental parameter. We foresee this approach will greatly increase the applicability of the learner across a diverse set of MS instrumentation. The capability developed here can be further extended by digital

and web-based mass spectrometric tools, like Global Natural Product Social (GNPS) molecular networking, which seeks to assist in identification and discovery of novel and previously uncharacterized molecules, including chemicals of concern.^{27,28}

The quality of MS data can also be enhanced by translating the current QQQ-based PIS methods to HRMS instrumentation. Specifically, we envision using pseudo precursor ion scans, which exploit high resolution mass spectrometers. The ultra-narrow m/z windows possible with this technique can greatly suppress signals that confound the nominal mass PIS data.

The present work serves as an initial effort highlighting the power of applying machine learning to alternative methods for MS identification of unknown or under-characterized toxins. Work is on-going to assess the influence of out-of-class MS data on model generation and fragment identification, the impact of open-source MS interpretation tools, and the application of high-resolution mass spectrometry towards enhanced identification of trichothecenes of concern to forensic, law enforcement, and public health communities..

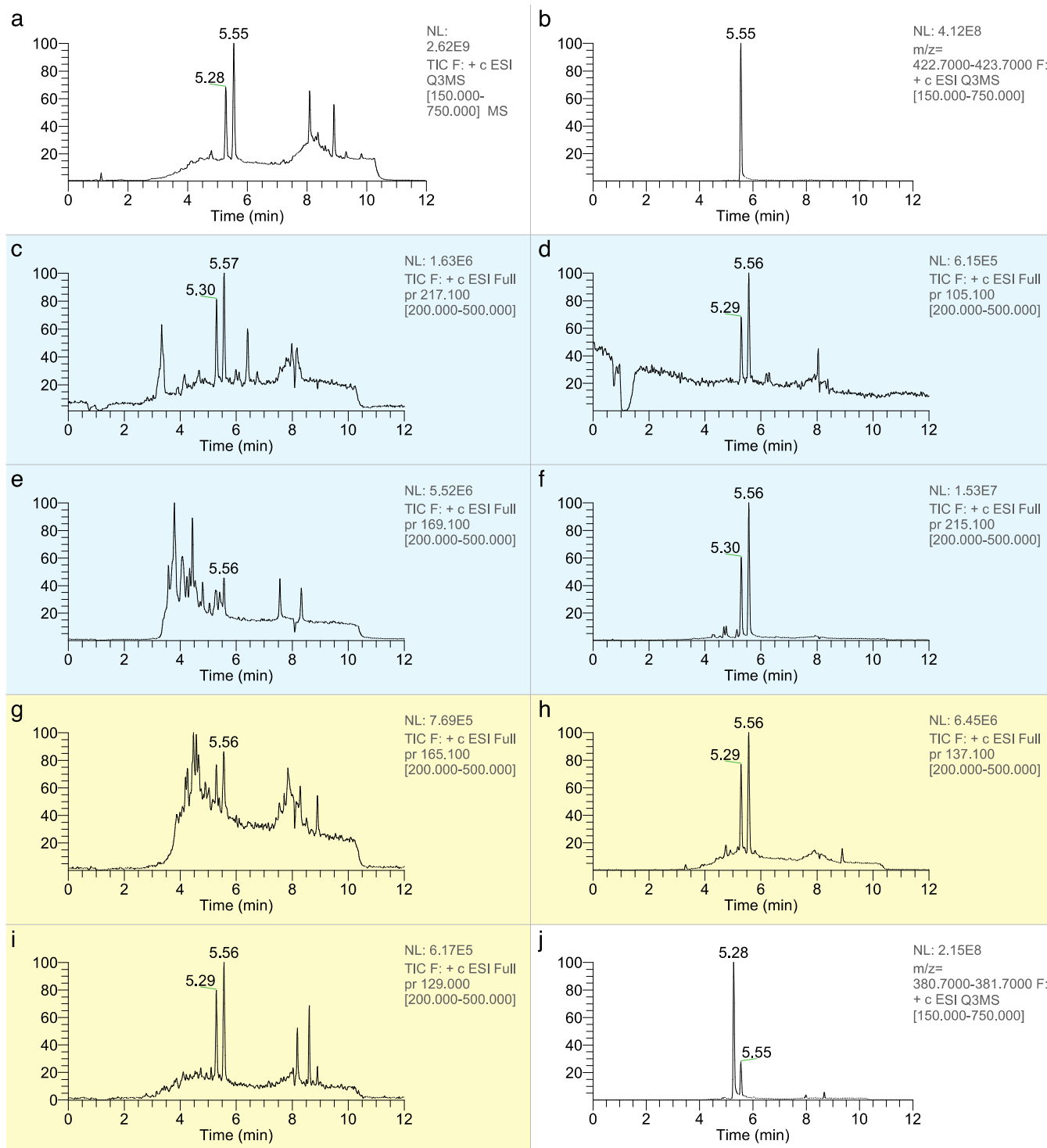


Figure 4. QQQ scans for a crude laboratory-synthesized trichothecene mixture of diacetyl-15-AcDON. (a) Total ion chromatogram of mixture; (b) extracted ion chromatogram (XIC) for $[M+H]^+$ of diacetyl-15-AcDON ($m/z = 423.2$); (c-f) Type A discriminative fragment PIS data for $m/z = 217.1, 105.1, 169.1,$ and 215.1 , respectively (in grey); (g-i) Type B discriminative fragment PIS data for $m/z = 165.1, 137.1,$ and 129.0 , respectively (in yellow); and (j) XIC for $[M+H]^+$ $m/z = 381.2$. Retention times are indicated for the two compounds of interest to the intended diacetyl-15-AcDON target.

ASSOCIATED CONTENT

Supporting Information

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

Extracted ion chromatograms of the ML-derived diagnostic fragments; synthetic details for tri-AcDON. (PDF)

AUTHOR INFORMATION

Corresponding Author

Brian Mayer - Forensic Science Center, Lawrence Livermore National Laboratory, 7000 East Ave. L-090, Livermore, CA, 94550. Orcid.org/0000-0003-1967-9802; Email: Mayer22@llnl.gov

Author Contributions

All authors have contributed equally. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

The Authors acknowledge the financial support from the United States Department of Homeland Security (70RSAT18KPM000183). This manuscript has been authored by Lawrence Livermore National Security, LLC under Contract No. DE-AC52-07NA2 7344 with the US. Department of Energy. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

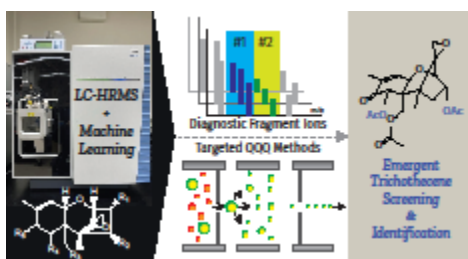
- (1) Etzel, R. A. Mycotoxins *JAMA- J. Am. Med. Assoc.* **2002**, *287* (4), 425-427.
- (2) McCormick, S. P.; Stanley, A. M.; Stover, N. A.; and Alexander, N. Trichothecenes: from simple to complex mycotoxins. *J. Toxins* **2011**, *3* (7), 802-814.
- (3) Moretti, A.; Susca, A.; Mule, G.; Logrieco, A. F.; Proctor, R. H. Molecular biodiversity of mycotoxigenic fungi that threaten food safety. *Int. J. Food Microbiol.* **2013**, *167* (1), 57-66.
- (4) Ran, R.; Wang, C.; Han, Z.; Wu, A.; Zhang, D.; Shi, J. Determination of deoxynivalenol (DON) and its derivatives: current status of analytical methods. *Food Control* **2013**, *34* (1), 138-148.
- (5) Liu, Z.-Y.; Yu, C.-H.; Wan, L.; Sun, Z.-L. Fragmentation study of five trichothecenes using electrospray hybrid ion trap/time-of-flight mass spectrometry with accurate mass measurements. *Int. J. Mass Spectrom.* **2012**, *309*, 133-140.
- (6) Freire, L.; Sant'Ana, A. S. Modified mycotoxins: an updated review on their formation, detection, occurrence, and toxic effects. *Food Chem. Toxicol.* **2018**, *111*, 189-205.
- (7) González-Jartín, J. M.; Alfonso, A.; Sainz, M. J.; Vieytes, M. R.; Botana, L. M. Detection of new emerging type-A trichothecenes by untargeted mass spectrometry. *Talanta* **2018**, *178*, 37-42.
- (8) Swanson, S. P.; Nicoletti, J.; Hood, H. D. Jr.; Buck, W. B.; Cote, L. M.; Yoshizawa, T. Metabolism of three trichothecene mycotoxins, T-2 toxin, diacetoxyscirpenol, and deoxynivalenol, by bovine rumen microorganisms. *J. Chromatogr.* **1987**, *414* (2), 335-342.
- (9) Strupat, K.; Scheibner, O.; Brominski, M. High-resolution, accurate-mass Orbitrap mass spectrometry - definitions, opportunities and advantages. Technical Note 64287, Thermo Scientific, 2016.

- (10) Lattanzio, V. M. T.; Ciasca, B.; Terzi, V.; Ghizzoni, R.; McCormick, S. P.; Pascale, M. Study of the natural occurrence of T-2 and HT-2 toxins and their glucosyl derivatives from field barley to malt by high-resolution Orbitrap mass spectrometry. *Food Addit. Contam. A* **2015**, *32* (10), 1647-1655.
- (11) Jia, W.; Shi, L.; Zhang, F.; Fan, C.; Chang, J.; Chu, X. Multiplexing data independent untargeted workflows for mycotoxins screening on a quadrupole-Orbitrap high resolution mass spectrometry platform. *Food Chem.* **2019**, *278*, 67-76.
- (12) Castellaneta, A.; Losito, I.; Cisternino, G.; Leoni, B.; Santamaria, P.; Calvano, C. D.; Bianco, G.; Cataldi, T. R. I. All ion fragmentation analysis enhances the untargeted profiling of glucosinolates in Brassica microgreens by liquid chromatography and high-resolution mass spectrometry. *J. Am. Soc. Mass Spectr.* **2022**, *33*, 2108-2119.
- (13) Nguyen, D. H.; Nguyen, C. H.; Mamitsuka, H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief. Bioinform.* **2018**, *20* (6), 2028-2043.
- (14) Van der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *P. Natl. A Sci.* **2016**, *113* (48), 13738-13743.
- (15) Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. MS2Analyzer: A Software for Small Molecule Substructure Annotations from Accurate Tandem Mass Spectra. *Anal. Chem.* **2014**, *86* (21), 10724-10731.
- (16) Mayer, B. P.; DeHope, A. J.; Mew, D. A.; Spackman, P. E.; Williams, A. M. Chemical Attribution of Fentanyl Using Multivariate Statistical Analysis of Orthogonal Mass Spectral Data. *Anal. Chem.* **2016**, *88* (8), 4303-4310.
- (17) Mayer, B. P.; Valdez, C. A.; DeHope, A. J.; Spackman, P. E.; Williams, A. M. Statistical Analysis of the Chemical Attribution Signatures of 3-Methylfentanyl and Its Methods of Production. *Talanta* **2018**, *186*, 645-654.
- (18) Williams, A. M.; Vu, A. K.; Mayer, B. P.; Hok, S.; Valdez, C. A.; Alcaraz, A. Part 3: Solid phase extraction of Russian VX and its chemical attribution signatures in food matrices and their detection by GC-MS and LC-MS. *Talanta* **2018**, *186*, 607-614
- (19) The R Project for Statistical Computing, <https://www.r-project.org/> (accessed 2023-02-21).
- (20) Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Benesty, M.; Lescarbeau, R.; Ziem, A.; Scrucca, L.; Tang, Y.; Candan, C.; Hunt, T. caret: Classification and Regression Training. R package version 6.0-76. 2018, <https://CRAN.R-project.org/package=caret> (accessed 2023-02-21)
- (21) Wei, T.; Simko, V. corrplot: Visualization of a Correlation Matrix. R package version 0.77. 2016, <https://CRAN.R-project.org/package=corrplot> (accessed 2023-02-21)
- (22) Venables, W. N.; Ripley, B. D. Modern Applied Statistics with S, 4th ed.; Springer, 2002.
- (23) Revolution Analytics; Weston, S. doMC: Foreach Parallel Adaptor for 'parallel'. R package version 1.3.4. 2015, <https://CRAN.R-project.org/package=doMC> (accessed 2023-02-21)
- (24) Kursu, M. B.; Rudnicki, W. R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36* (11), 1-13.
- (25) Hastie, T.; Tibshirani, R.; Friedman, J. Boosting and Additive Trees. In *The Elements of Statistical Learning*, 2nd ed.; Springer-Verlag, 2008; pp 337-388.
- (26) Shi, L.; Westerhuis, J. A.; Rosén, J.; Landberg, R.; Brunius, C. Variable selection and validation in multivariate modelling. *Bioinform.* **2019**, *35* (6), 972-980.
- (27) Aron, A. T.; Gentry, E. C.; McPhail, K. L.; Nothias, L.-F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J. M.; Sikora, N.; Vargas, F.; van der Hooft, J. J.; Ernst, M.; Kang, K. B.; Aceves, C. M.; Caraballo-Rodríguez, A. M.; Koester, I.; Weldon, K. C.; Bertrand, S.; Roullier, C.; Sun, K.; Tehan, R. M.; Boya P., C. A.

Christian, M. H.; Gutiérrez, M.; Ulloa, A. M.; Tejada Mora, J. A.; Mojica-Flores, R.; Lakey-Beitia, J.; Vázquez-Chaves, V.; Zhang, Y.; Calderón, A. I.; Tayler, N.; Keyzers, R. A.; Tugizimana, F.; Ndlovu, N.; Aksenov, A. A.; Jarmusch, A. K.; Schmid, R.; Truman, A. W.; Bandeira, N.; Wang, M.; Dorrestein, P. C. Reproducible Molecular Networking of Untargeted Mass Spectrometry Data Using GNPS. *Nature Protocols* **2020**, *15* (6), 1954–1991.

(28) Vincenti, F.; Montesano, C.; Di Ottavio, F.; Gregori, A.; Compagnone, D.; Sergi, M.; Dorrestein, P. Molecular Networking: A Useful Tool for the Identification of New Psychoactive Substances in Seizures by LC–HRMS. *Frontiers in Chemistry* **2020**, *8*, 572952.

Table of Contents artwork



- 1 Etzel, R. A. *JAMA-J. Am. Med. Assoc.* **2002**, *287* (4), 425-427.
- 2 McCormick, S. P.; Stanley, A. M.; Stover, N. A.; and Alexander, N. J. *Toxins* **2011**, *3* (7), 802-814.
- 3 Moretti, A.; Susca, A.; Mule, G.; Logrieco, A. F.; Proctor, R. H. *Int. J. Food Microbiol.* **2013**, *167* (1), 57-66.
- 4 Ran, R.; Wang, C.; Han, Z.; Wu, A.; Zhang, D.; Shi, J. *Food Control* **2013**, *34* (1), 138-148.
- 5 Liu, Z.-Y.; Yu, C.-H.; Wan, L.; Sun, Z.-L. *Int. J. Mass Spectrom.* **2012**, *309*, 133-140.
- 6 Freire, L.; Sant'Ana, A. S. *Food Chem. Toxicol.* **2018**, *111*, 189-205.
- 7 González-Jartín, J. M.; Alfonso, A.; Sainz, M. J.; Vieytes, M. R.; Botana, L. M. *Talanta* **2018**, *178*, 37-42.
- 8 Swanson, S. P.; Nicoletti, J.; Hood, H. D. Jr.; Buck, W. B.; Cote, L. M.; Yoshizawa, T. *J. Chromatogr.* **1987**, *414* (2), 335-342.
- 9 Strupat, K.; Scheibner, O.; Brominski, M. High-resolution, accurate-mass Orbitrap mass spectrometry - definitions, opportunities and advantages. Technical Note 64287, Thermo Scientific, 2016.
- 10 Lattanzio, V. M. T.; Ciasca, B.; Terzi, V.; Ghizzoni, R.; McCormick, S. P.; Pascale, M. *Food Addit. Contam. A* **2015**, *32* (10), 1647-1655.
- 11 Jia, W.; Shi, L.; Zhang, F.; Fan, C.; Chang, J.; Chu, X. *Food Chem.* **2019**, *278*, 67-76.
- 12 Castellana, A.; Losito, I.; Cisternino, G.; Leoni, B.; Santamaria, P.; Calvano, C. D.; Bianco, G.; Cataldi, T. R. I. All ion fragmentation analysis enhances the untargeted profiling of glucosinolates in *Brassica* microgreens by liquid chromatography and high-resolution mass spectrometry. *J. Am. Soc. Mass Spectr.* **2022**, *33*, 2108-2119.
- 13 Nguyen, D. H.; Nguyen, C. H.; Mamitsuka, H. *Brief. Bioinform.* **2018**, *20* (6), 2028-2043.
- 14 Van der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S. P. *Natl. A Sci.* **2016**, *113* (48), 13738-13743. 3
- 15 Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O. *Anal. Chem.* **2014**, *86* (21), 10724-10731e
- 16 Mayer, B. P.; DeHope, A. J.; Mew, D. A.; Spackman, P. E.; Williams, A. M. *Anal. Chem.* **2016**, *88* (8), 4303-4310. 4
- 17 Mayer, B. P.; Valdez, C. A.; DeHope, A. J.; Spackman, P. E.; Williams, A. M. *Talanta* **2018**, *186*, 645-654. 6
- 18 Williams, A. M.; Vu, A. K.; Mayer, B. P.; Hok, S.; Valdez, C. A.; Alcaraz, A. *Talanta* **2018**, *186*, 607-614
- 19 The R Project for Statistical Computing, <https://www.r-project.org/> (accessed 2023-02-21).
- 20 Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Benesty, M.; Lescarbeau, R.; Ziem, A.; Scrucca, L.; Tang, Y.; Candan, C.; Hunt, T. caret: Classification and Regression Training. R package version 6.0-76. 2018, <https://CRAN.R-project.org/package=caret> (accessed 2023-02-21)
- 21 Wei, T.; Simko, V. corrplot: Visualization of a Correlation Matrix. R package version 0.77. 2016, <https://CRAN.R-project.org/package=corrplot> (accessed 2023-02-21)
- 22 Venables, W. N.; Ripley, B. D. Modern Applied Statistics with S, 4th ed.; Springer, 2002.
- 23 Revolution Analytics; Weston, S. doMC: Foreach Parallel Adaptor for 'parallel'. R package version 1.3.4. 2015, <https://CRAN.R-project.org/package=doMC> (accessed 2023-02-21)
- 24 Kursat, M. B.; Rudnicki, W. R. J. Stat. Softw. 2010, *36* (11), 1-13.
- 25 Hastie, T.; Tibshirani, R.; Friedman, J. Boosting and Additive Trees. In *The Elements of Statistical Learning*, 2nd ed.; Springer-Verlag, 2008; pp 337-388.
- 26 Shi, L.; Westerhuis, J. A.; Rosén, J.; Landberg, R.; Brunius, C. *Bioinform.* **2019**, *35* (6), 972-980.
- 27 Aron, A. T.; Gentry, E. C.; McPhail, K. L.; Nothias, L.-F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J. M.; Sikora, N.; Vargas, F.; van der Hooft, J. J.; Ernst, M.; Kang, K. B.; Aceves, C. M.; Caraballo-Rodríguez, A. M.; Koester, I.; Weldon, K. C.; Bertrand, S.; Roullier, C.; Sun, K.; Tehan, R. M.; Boya P., C. A.; Christian, M. H.; Gutiérrez, M.; Ulloa, A. M.; Tejada Mora, J. A.; Mojica-Flores, R.; Lakey-Beitia, J.; Vásquez-Chaves, V.; Zhang, Y.; Calderón, A. I.; Tayler, N.; Keyzers, R. A.; Tugizimana, F.; Ndlovu, N.; Aksenov, A. A.; Jarmusch, A. K.; Schmid, R.; Truman, A. W.; Bandeira, N.; Wang, M.; Dorrestein, P. C. Reproducible Molecular Networking of Untargeted Mass Spectrometry Data Using GNPS. *Nature Protocols* **2020**, *15* (6), 1954-1991.
- 28 Vincenti, F.; Montesano, C.; Di Ottavio, F.; Gregori, A.; Compagnone, D.; Sergi, M.; Dorrestein, P. Molecular Networking: A Useful Tool for the Identification of New Psychoactive Substances in Seizures by LC-HRMS. *Frontiers in Chemistry* **2020**, *8*, 572952.