Analytica Chimica Acta 1112 (2020) 34–45

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Using isotopic envelopes and neural decision tree-based *in silico* fractionation for biomolecule classification

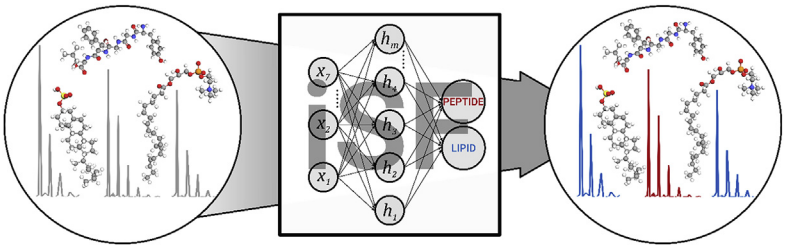
Luke T. Richardson , Matthew R. Brantley , Touradj Solouki *

Department of Chemistry and Biochemistry, 76706, 101 Bagby Ave., Baylor University, Waco, TX, USA

HIGHLIGHTS

- A new neural network-based approach for classification of proteomics, lipidomics, and other omics MS data is introduced.
- The *iSF* approach utilizes a neural decision tree (NDR) to sequentially classify biomolecules.
- The *iSF* classifications are based on the experimentally acquired MS-data and isotopic patterns.
- NDR with supervised binary & multi-target classifiers allows for classification of polypeptides & lipid subcategories.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 May 2019
Received in revised form
16 January 2020
Accepted 17 February 2020
Available online 20 February 2020

Keywords:
Mass spectrometry
Chemometrics
Feedforward neural network
Neural decision tree
Isotopic envelope

ABSTRACT

Untargeted mass spectrometry (MS) workflows are more suitable than targeted workflows for high throughput characterization of complex biological samples. However, analysis workflows for untargeted methods are inadequate for characterization of complex samples that contain multiple classes of compounds as each chemical class might require a different type of data processing approach. To increase the feasibility of analyzing MS data for multi-class/component complex mixtures (*i.e.*, mixtures containing more than one major class of biomolecules), we developed a neural network-based approach for classification of MS data. In our *in silico* fractionation (iSF) approach, we utilize a neural decision tree to sequentially classify biomolecules based on their MS-detected isotopic patterns. In the presented demonstration, the neural decision tree consisted of two supervised binary classifiers to positively classify polypeptides and lipids, respectively, and a third supervised network was trained to classify lipids into the eight main sub-categories of lipids. The two binary classifiers assigned polypeptide and lipid experimental components with 100% sensitivity and 100% specificity; however, the 8-target classifier assigned lipids into their respective subclasses with 95% sensitivity and 99% specificity. Here, we discuss important relationships between class-specific chemical properties and MS isotopic envelopes that enable analyte classification. Moreover, we evaluate the performance characteristics of the utilized networks.

1. Introduction

Mass spectrometry (MS)-based profiling strategies have been developed to handle increased throughput [1–3] and sample

Abbreviations: iSF, *in silico* fractionation; FFNN, feedforward neural network; NDT, neural decision tree; KMD, Kendrick mass defect; *m/z*, mass-to-charge ratio; NA, natural abundance; MRP, mass resolving power; TPR, true positive rate; PPV, positive predictive value; IM, ion mobility; T, target class.

* Corresponding author. Department of Chemistry and Biochemistry, Baylor University Sciences Building, One Bear Place #97348, Waco, TX, 76798, USA.
E-mail address: Touradj_Solouki@baylor.edu (T. Solouki).

https://doi.org/10.1016/j.aca.2020.02.036
0003-2670/

complexity [4–6]. Combined analysis of multiple classes of biomolecule (*i.e.*, integrated/multiomics) can be advantageous and enable evaluation of correlations between different, but related, biological systems (*e.g.*, the proteome and lipidome) [7–9]. Conventionally, each class of compounds in biological samples is interrogated separately and then the results from multiple MS analyses are integrated. However, these approaches are not ideal for high throughput workflows as they often require time-intensive, off-line (as opposed to "in-line" with MS injection) sample fractionation methods. Off-line fractionation methods reduce sample complexity, prior to MS analysis, by isolating classes or groups of compounds based on their polarity, size, charge, or other physicochemical properties. In conventional analysis of biological samples, off-line sample fractionation allows researchers to make reasonable assumptions about classes of analytes to be detected and, hence, determine the most appropriate MS operational modes and sample-specific post-acquisition data analysis workflows.

In contrast to conventional class-specific characterization of multi-component samples, simultaneous characterization of these complex samples can increase throughput and reduce biases associated with targeted sample fractionation techniques. Currently, combined sampling is complicated by biases (a) inherent to ionization methods (*e.g.*, basicity and proton transfer kinetics [10], solvent types and ionization efficiency/proton affinity differences based on analyte polarity in electrospray ionization (ESI) [11], and matrix dependent ionization efficiencies of different classes of molecules in matrix assisted laser desorption ionization (MALDI) [12–15]) and (b) lack of in-line physical separation and sample preparation methods compatible with multi-class analyses in the literature. Recently introduced "integrated" liquid-, gas-, and solid-phase ion source technologies seek to eliminate ionization biases and accommodate samples of greater biological diversity [5,6,16–24]. Numerous MS-based techniques, focused on the increasingly popular field of multi-omics analyses [25–29], could directly benefit from the presented *in silico* fractionation (iSF) approach herein which employs a post-data acquisition tactic for analyte classifications of multi-class component samples. Despite matrix dependent ionization biases for different classes of biomolecule exhibited by MALDI matrixes, MALDI surface sampling coupled with ion mobility (IM)-MS has been used for simultaneous analysis of multi-class mixtures of small biomolecules directly from tissue [30–32]; iSF could be applied to such MALDI generated complex data sets. Moreover, a preliminary demonstration of Omni-MS showed a sample preparation and separation strategy for concurrent LC-MS analysis of electrolytes, small molecules, lipids, polypeptides, nucleic acids, and polysaccharides [33]. We expect that technologies for simultaneous multi-omics analyses will continue to develop in pursuit of higher throughput and information density per data acquisition. However, downstream analysis of such multi-class acquisitions will remain a challenge given that: (a) assumptions regarding the class of detected ions may not be reliable and (b) different classes may have different requisite MS operational modes and/or analytical workflows. Thus, multi-omics MS analyses require some method for discriminating between (and identifying) classes of analytes. We have previously shown that MS isotopic envelope of molecular ion signals contain sufficient information to discriminate between compounds containing different functional groups or specific elements [34]; here, we aim to demonstrate the strength of combining neural networks and isotopic pattern-based biomolecule class designation for multi-omics characterization of multi-class sample mixtures.

The isotopic envelope (*i.e.*, *m/z* range containing all isotopologues of a specific compound) is a readily observed feature in high resolution mass spectra [35]. In theory, each elemental composition has a unique MS isotopic fine structure [36] that is dependent on the mass contributions of each element and relative abundances of heavy isotopes. Given that chemical homology is conserved (often to different degrees) within classes of biomolecules, the elemental compositions, and thus the MS isotopic envelopes, of biomolecules reflect this similarity. This idea of intra-class chemical homology is at least tacitly understood in untargeted MS experiments that utilize a combination of sub-ppm error (exact mass measurement assignments for presumably resolved peaks) and isotopic envelope information to generate compound elemental compositions and cluster analytes roughly by class in van Krevelen visualization — though without means for definitive classification) [37,38]. However, high (H) and ultrahigh (UH) resolution MS instruments often lack the necessary mass resolving power (MRP) and mass measurement accuracy to consistently yield accurate elemental compositions for large molecules (molecular weight (MW) > 500 Da) containing elements beyond carbon, hydrogen, nitrogen, and oxygen [37,39]. Moreover, UH resolution FT-MS instruments (MRP > 250,000) are unfavorable for high throughput applications and high scan rate rate instruments (e.g., time-of-flight (TOF) instruments with ~10,000–100,000 MRP) are preferred. *A priori* knowledge of analyte class is often necessary to determine the elements allowed for use in the generation of elemental compositions in order to shorten the often-lengthy list of possible elemental compositions produced within a margin of experimental error [40]. Without *a priori* knowledge, determination of the elemental composition becomes exponentially more difficult due to the possibility of having myriad different elemental compositions that can yield the measured mass and isotopic envelope [41]. As expected, the length of the produced "list" of elemental compositions generally decreases with increased MRP, resulting in higher confidence compound identifications and/or classifications; however, it is often impossible to reduce the number of possibilities to a single, high confidence determination [41]. Elemental composition, and therefore compound class, could theoretically be determined via computationally intensive solutions to the polynomial model that describes the summation of elemental contributions [42] to isotopologue peaks given error-free conditions. In general, the effect of non-ideal experimental conditions on exact mass determinations of elemental composition and inapplicability of UHMRP MS instrumentation in high throughput applications necessitate an automated approach for definitive analyte classification in multi-omics analyses that is less dependent on MRP through utilization of other features in the isotopic envelope.

Feedforward neural networks (FFNNs, described by Bishop [43]) offer a potential solution to this problem. Trained FFNNs excel in estimating non-linear, high dimensional relationships to discern hidden patterns and classify network inputs. FFNNs are based on the early perceptron model in which network inputs are recursively weighted, summed, and submitted to an activation function; in this simple scenario, training stops once the output of an activation function exceeds a specified value, producing a linear function that serves as a binary classifier [44]. FFNNs utilize an expanded hidden perceptron layer architecture to more effectively map complex systems and solve problems with high dimensionality (*i.e.*, those with large sets of inputs and multiple output functions) [45]. We hypothesized that FFNNs could effectively estimate the non-linear relationships that describe the features of the MS isotopic envelope to discriminate between classes of biomolecules in MS data (without the painstaking task of mathematically accounting for the discrete contributions of each element). To test our hypothesis, we examined both theoretical and experimentally acquired MS data. Additionally, given a sufficiently large and varied training set and an appropriate network architecture, we hypothesized that FFNNs

could be sufficiently generalized such that experimental error and noise sources present in mass spectral data would not pose a serious concern [46]. In this paper, we confirm the effectiveness of FFNNs for biomolecule class identification in multi-omics analyses and its tolerance for handling experimental measurement errors, common in HMRP MS data, that can impede compound identification, elemental composition determination, and/or classification by exact mass measurement.

Specifically, we present an FFNN-based chemometric analysis tool for classification of small biomolecules (*e.g.*, lipids, polypeptides, nucleic acids, and glycans) by utilizing their commonly acquired soft ionization MS data produced from TOF instrumentation. We show that a series of FFNNs can be trained using features extracted from the centroided isotopic envelopes (*i.e.*, the mass-to-charge (*m/z)* and relative isotopologue peak intensities/ratios) to classify individual sample components into a selection of biomolecular class targets. Using these trained FFNNs, we show that a neural decision tree (NDT) [47] can be constructed to utilize MS data sets and sequentially classify analytes in a representative multi-class component sample containing polypeptides, metabolites, and lipids. Moreover, we show that classified lipids can be assigned to their respective subclasses and demonstrate how this technique can provide simple, qualitative results that can be interpreted by non-MS experts. In addition to classification of theoretically generated MS data, we confirm the validity of *in silico* fractionation (iSF) by using experimental data and successfully identifying its components. We also show that network classification operations can be rapid and comparable to time-of-flight (TOF) MS detection event time scales [48]. We provide an analysis and a discussion of the chemical basis for the *in silico* fractionation (iSF) technique and its adaptability to various types of MS methodologies.

2. Materials and methods

2.1. Neural network input preprocessing

To generate theoretical *m/z* values and isotopic patterns for selected classes of biomolecules (*viz.*, lipids, glycans, non-lipid metabolites, polypeptides, DNA, and RNA), elemental compositions of a total of 36,973 molecules were acquired from four different sources as indicated in the following text. Elemental compositions for these 36,973 molecules were sourced from (1) the LIPID MAPS Structural Database (lipids, n = 7473) [49], (2) the Consortium of Functional Glycomics (glycans, n = 1750), and (3) the Human Metabolome Database (organic, non-lipid metabolites, n = 7454) [50]. Lastly, polypeptide (n = 7465), deoxyribonucleic acid (DNA, n = 7140), and ribonucleic acid (RNA, n = 5691) elemental compositions were generated by (4) an in-house written python (CPython 3.6.2; Python Software Foundation, DE) script that pseudo-randomly generated heteropolymer sequences corresponding to expected ESI charge states based on biopolymer chain lengths [51] and converted them to elemental compositions (see Appendix Script A.1). For example, polypeptides in the +1 charge state result from a gaussian-like distribution of polymer chain lengths between about 5 and 17 residues as demonstrated by Xie et al. [51]. Therefore, the chain lengths for singly-charged, protonated polypeptides were generated by a gaussian probability function with bounds to match the experimentally observed distribution. The centroided isotopic distribution profiles for some protonated, deprotonated, and sodiated ESI adducts and charge states (respective to the class of biomolecule) were calculated in the enviPat 2.2 package [52] in R (ver. 3.4.3; R Foundation for Statistical Computing). The molecular adduct forms and charge states included in the training set were not inclusive of all commonly observed types; the training set was limited in this regard due to deteriorating network training performance as the inclusion of multiple forms presumably added extraneous dimensionality to the FFNN training set. Isotopic profiles were generated considering respective physical realities of ion production in ESI for each class as follows: polypeptides as protonated ions ranging from $[M+H]^+$ to $[M+6H]^{6+}$, glycans as positively charged, sodiated ions $[M+Na]^+$ and $[M+2Na]^{2+}$ and as negatively charged, deprotonated ions ranging from $[M-H]^-$ to $[M-3H]^{3-}$, lipids as positively charged, protonated ions $[M+H]^+$, and oligonucleotides as negatively charged, deprotonated ions ranging from $[M-H]^-$ to $[M-6H]^{6-}$.

2.2. Neural network training

FFNNs were constructed with a custom script (see Appendix Script A.2) written in M (MATLAB R2018a; The MathWorks Inc., Natick, MA). Three FFNNs, denoted as PEPNET, LIPNET, and LIPSUBNET, were used in this study. PEPNET was a binary classifier to assign network inputs into the peptide (T_P) or non-peptide (T_{NP}) target class. Whereas LIPNET was a binary classifier to assign network inputs into the lipid (T_L) or non-lipid (T_{NL}) target class; for lipid subclass categorization, LIPSUBNET (an 8-target classifier) was used to assign previously classified T_L inputs into class targets corresponding to lipid subclasses. Designated lipid subclasses in LIPSUBNET included: fatty acyls (T_{FA}), glycerolipids (T_{GL}), glycerophospholipids (T_{GP}), polyketides (T_{PK}), prenol lipids (T_{PR}), saccharolipids (T_{SL}), sphingolipids (T_{SP}), and sterol lipids (T_{ST}).

The hidden layer architecture (i.e., number of hidden layers and number of nodes per hidden layer) for each type of FFNN was chosen such that classification accuracy for the input data set used for training would be maximized. For FFNNs with more than one hidden layer, the number of nodes per layer were kept the same across layers as networks already exhibited high performance and further optimization was unnecessary. Training times were short enough that they were not considered in determining network architecture. 10 FFNNs were trained for each tested network architecture, and performance was evaluated by the mean percent error (percent of false positive and false negative classifications over all classifications). Network architectures were chosen for the minimization of mean percent error (for more information regarding network architecture optimization, see Appendix Figures B1 and B2). When training the PEPNET and LIPNET binary classifiers, it became apparent that small changes in hidden layer sizes had notable effects on performance; therefore, hidden layer neurons were increased by 10 during optimization tests. The optimized hidden layer architecture of LIPNET and PEPNET was constituted by 2 layers with 30 perceptrons each. For the 8-target classifier (LIPSUBNET), network performance was relatively insensitive to small changes in hidden layer neurons; thus, hidden layer neurons were increased by the power function 2^n to evaluate a large range of hidden layer neurons. The optimized hidden layer architecture was constituted by 2 layers with 64 perceptrons each. Training bias due to uneven class representation was accounted for by using the "growth method" (Brantley et al. [53]) such that each class had equal numbers of representations in the dataset. The input data set was then randomly divided (*i.e.*, MATLAB function "dividerand") for network training as follows: 70% in the training set, 15% in the validation set, and 15% in the test set. Networks were trained using scaled conjugate gradient backpropagation (*i.e.*, MATLAB function "trainscg") with an early stopping method [54] to avoid overfitting. Performance during training was monitored using cross-entropy as an error metric (MATLAB function "crossentropy"). Network inputs were 7 numbers (henceforth referred to as "input vectors") in the following order: (1) the exact mass monoisotopic *m/z* value to four decimal places (to be consistent

with TOF data utilized in this demonstration) corresponding to the first isotopic peak or "A" (where "A" notation is based on the designation introduced by McLafferty et al. [35]), (2) the intensity of "A" normalized to the most abundant isotopologue (i.e., normalized to the highest peak within the isotopic envelope for the molecular ion designated as 100% relative abundance), (3) the relative intensity for the second isotopologue or A+1, (4) the relative intensity for the third isotopologue or A+2, (5) the ratio of A/A+1, (6) the ratio of A+1/A+2, and (7) the Kendrick mass defect [55] (KMD, relative to CH_2 integer mass scale) of the monoisotopic *m/z*. The output layers were limited to two (for PEPNET and LIPNET) or eight (for LIPSUBNET) nodes corresponding to the classification targets of each neural network. The PEPNET output layer nodes correspond to: 1) polypeptide and 2) non-polypeptide. The LIPNET output layer nodes correspond to: 1) lipid and 2) non-lipid. The LIPSUBNET output layer nodes correspond to: 1) fatty acyl, 2) glycerolipid, 3) glycerophospholipid, 4) polyketide, 5) prenol lipid, 6) saccharolipid, 7) sphingolipid, and 8) sterol lipid.

Supervised training target outputs were provided such that the presence of a class of biological molecules was indicated by a value of 1.0 and absence by a value of 0.0. The same network inputs were used for each FFNN, but the supervised training target outputs were changed depending on FFNN being trained (e.g., in PEPNET, polypeptide components had target outputs of 1.0 and lipid/metabolite/glycan/DNA components had target outputs of 0.0; the same components were used in LIPNET, but polypeptide component target outputs were changed to 0.0, and lipid component target outputs were changed to 1.0). Additionally, LIPSUBNET was trained with only lipid components. For each scored component (in either the training set or an experimental test set), the maximum predicted value from the set of values corresponding with each class predicted by the network is taken as the predicted class. Ten FFNNs were trained for each type of network (i.e., PEPNET, LIPNET, and LIPSUBNET) with the optimal hidden layer architectures, and the network with the minimum mean percent error across the training, validation, and test datasets was selected as the best performing network for further analyses. All networks were trained using a desktop computer (OptiPlex 7050 Tower; Dell Computer, Round Rock, TX) equipped with a 4-core processor (Intel® Core™ i5-7500) and 16 GB of RAM.

2.3. Mass spectrometry methods and multi-class component test set generation

To test the iSF approach on experimental data, a multi-class component test set was generated. *Firstly*, an LC-MS output matrix (rows and columns corresponding to LC scan numbers and *m/z* axis data points, respectively) from a rat brain tryptic protein digest analysis (collected in-house for *m/z* range 50–2000; see below for MS methodology) was added to LC-MS output matrix (also, 50–2000 *m/z* range) from a lipidomics LC-MS output. The lipidomics data was downloaded from the Chorus Project (Stratus Biosciences, Seattle, WA) mass spectrometry file sharing database [56]. Available lipidomics data from Chorus [56] included a limited range of LC elution time and hence, only scans corresponding to ~15 min (scan numbers up to ~900, at an acquisition rate of 1 Hz for both lipidomics and proteomics) was included for further analysis (e.g., Fig. 5 shows analysis results for scan range of ~400–900). Matrix addition was performed by the MassLynx (V4.1, Waters, Milford, MA) "Combine all functions" tool. *Secondly*, LC eluting components were found in the combined data set and manually (on a random order) were selected for inclusion in the test set. The data features necessary for the neural network were exported from the MassLynx mass spectrum data viewer. *Thirdly*, only the eluting components that could be identified were included so that their true chemical classes would be known for the evaluation of FFNN. Components that were identified as peptides were sequenced and identified in ProteinLynx Global Server (PLGS, Waters Corp., Milford, MA); the LC retention time recorded by PLGS was confirmed for each peptide. Components that were identified as lipids and metabolites were identified in Progenesis QI (Nonlinear Dynamics, Durham, NC) by exact mass matching (<5 ppm mass measurement error) in the LIPID MAPS and Metlin databases, respectively. Because it was necessary to confirm to which group (lipidomics or proteomics data set) each component belonged, identified lipid components were confirmed to be absent in the original proteomics data set. Likewise, identified peptide components were confirmed to be absent in the original lipidomics data set.

The downloaded LC-MS lipidomics data set was acquired using a Waters Xevo-G2 QToF (Waters Corp., Milford, MA) in positive-ion mode [56]. To collect the rat brain protein tryptic digest data, ultra-performance liquid chromatography (UPLC)-ion mobility-enhanced MS^E ($HDMS^E$) on a Synapt G2-S HDMS (Waters Corp., Milford, MA) operating in positive-ion mode was used. Pettit et al. previously described the procedure for rat brain tissue sample preparation [57]. A 10 mg rat brain cortex tissue section was homogenized via ultrasonication probe in lysis solution (100 mM ammonium bicarbonate, 1% (w/v%) sodium dodecyl sulfate, 10 mM tris(2-carboxyethyl) phosphine hydrochloride, and 40 mM 2-chloroacetamide; all chemicals from Fisher Scientific, NH). Bottom-up proteomics samples were prepared according to the filter-aided sample preparation protocol [58], which involves sequential wash and high mass (>3 kDa) filtration (MilliporeSigma, MA) that washes lipids and detergents from the sample. Samples were digested with sequencing-grade trypsin (Promega, Madison, WI) overnight at 37 °C. PLGS was employed for peptide identification of the tryptic protein digests. The lipidomics and proteomics MS data were both acquired at a mass resolution of ~22,000 (see Appendix Figure B3 for exemplary peptide and lipid isotopic envelopes from the combined dataset with mass resolution calculations).

HeLa digest peptides and a rat brain lipid extract were mixed and analyzed via UPLC-IM-MS on a Synapt G2-S HMDS operating in positive-ion mode. The HeLa digest peptides were purchased as a standard from Thermo Fisher Scientific (Waltham, MA). A 10 mg rat brain tissue section was homogenized via ultrasonication in ice-cold 0.1% ammonium acetate and prepared as described by Matyash et al. [59] The HeLa digest peptide mixture was reconstituted to a concentration of 60 ng/μL in 0.1% formic acid in water, and the lipid extract was reconstituted to a final volume of 100 μL in an isopropanol/acetonitrile/water (4:3:1, v/v/v) solution. Equivalent volumes of each mixture were mixed, and a volume of 5 μL was injected on column (150 ng of peptides and lipids extracted from 250 μg of rat brain tissue). The mixture was separated by reversed-phase chromatography on the NanoAquity UPLC using a Symmetry C18 trap column (5 μm, 180 μm × 20 mm, Waters Corp., Milford, MA) and a BEH130C18 analytical column (1.8 μm, 100 μm × 100 mm, Waters Corp., Milford, MA). Analytes were separated with a binary solvent gradient with 10 mM ammonium formate and 0.1% formic acid in water (mobile phase A) and isopropanol/acetonitrile (90:10, v/v) with 10 mM ammonium formate and 0.1% formic acid. The gradient ramped linearly from 5% to 99% B over 70 min with an isocratic hold at 99% B for 4 min at 0.4 μL/min. Capillary voltage was set at 2.7 kV, and the source temperature was set at 100 °C. Travelling wave ion mobility conditions were set at default settings. Analytes for classification were selected from three time periods of the chromatogram: peptides from 10 to 30 min, lipids from 55 to 80 min, and unknowns from 30 to 55 min. The time periods were selected based on an understanding peptide and lipid elution behavior in reversed-phase separations. The

10–30 min period roughly corresponded to 13–40% organic solvent composition, which is used for reversed-phase peptide separations [57]. The 55–80 min period roughly corresponded to 70–99% organic solvent composition, which is common for reversed-phase lipid separations [60]. Selected analytes were limited to those with monoisotopic *m/z* values between 500 and 1200 *m/z*. Before 10 min, only unretained components and background signal were observed, and, between 80 and 90 min, only background polymer signals were detected, and therefore these periods were excluded.

3. Results and Discussion

The following sections describe the generation and implementation of the iSF workflow; a neural decision tree-based method is introduced that utilizes MS data for the classification of small molecules. The physical and MS principles that constitute the basis for iSF as well as feature selections for neural network training are discussed in detail. Additionally, examples of iSF analyses of an experimental LC-MS datasets containing multiple biological classes of molecule are provided.

3.1. MS isotopic envelope feature selection

Classes of biomolecules, such as polypeptides and nucleic acids, consist of biopolymers that are composed of monomeric subunits and are limited in elemental diversity for a given polymer length by the number of possible monomers. Contrarily, biomolecules such as lipids are predominantly non-polymeric structures and include more diverse groups. Fig. 1 displays the histogram distributions for elemental compositions of nucleic acids (purple), glycans (green), polypeptides (blue), and lipids (red) as a function of their respective mass percent composition (w/w%) values for hydrogen (Fig. 1a), carbon (Fig. 1b), nitrogen (Fig. 1c), and oxygen (Fig. 1d). For example, lipids are hydrocarbon-rich and thus the mass percent compositions for carbon and hydrogen are in the ranges of ~35–90% (Fig. 1b) and ~3–14% (Fig. 1a), respectively. Additionally, lipids are substituted with a wide variety of polar functional groups and hence they are composed of ~0–14% nitrogen (w/w%, Fig. 1c) and ~4–52% oxygen (w/w%, Fig. 1d), respectively. Conversely, due to the constrained modes of elongation and incorporation of a limited number of possible monomers into their structures, polypeptides, nucleic acids, and glycans have narrower distributions of carbon (w/w%) and nitrogen (w/w%) contents; additionally, the distribution of oxygen composition in glycans is also very confined (green histogram in Fig. 1d). In fact, both nucleic acids and glycans have two nearly non-overlapping elemental percent distributions that are away from other distributions (e.g. for nucleic acids, hydrogen (purple histogram in Fig. 1a) and carbon (purple histogram in Fig. 1b) and, for glycans, carbon (green histogram in Fig. 1b) and oxygen (green histogram in Fig. 1d)). The carbon w/w% of polypeptides range from ~41 to 62% (blue histogram in Fig. 1a) and, though slightly overlap with that of glycans; this range is completely distinguished from the range of carbon w/w% for nucleic acids (~37–39%). Interestingly, an analytical tool for classification tasks in UHMRP MS data, called the van Krevelen diagram, visualizes and groups detected compounds using elemental molar ratios (e.g., the ratio of H/C and O/C atoms) derived from their calculated elemental compositions [38]. Given that elemental molar compositions and mass compositions are related with elemental molar mass, classifications made by van Krevelen diagrams inherently utilize a similar approach to describe class distributions (Fig. 1). Given that van Krevelen diagram visualization requires ultrahigh mass resolving power FT-MS instruments (generally MRP > 200,000) to generate high confidence chemical compositions from exact mass measurements of monoisotopologue

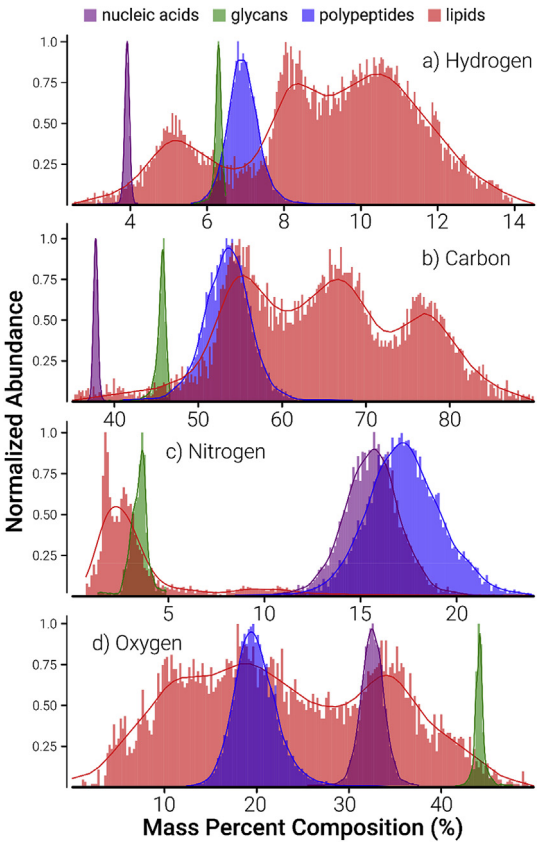


Fig. 1. Histograms of nucleic acid (purple), glycan (green), polypeptide (blue), and lipid (red) compositions as a function of mass percent composition of (a) hydrogen, (b) carbon, (c) nitrogen, and (d) oxygen overlaid with a kernel density estimation line plot. The nucleic acid, polypeptide, and glycan class distributions are generally gaussian-like and overlap minimally, except for those for nitrogen mass percent composition in which there is significant overlap. Lipid species are generally distributed over a wide range of elemental mass compositions. Elemental composition of a significant number of the selected lipids lacked nitrogen (~37%), and hence the nitrogen histogram for lipids (which includes contributions from 63% of all lipids) does not display contributions from those species that contained no nitrogen (i.e., 37% of the lipids used to generate (c) contained 0% nitrogen). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

peaks, the technique is generally unviable for conventional HMRP instruments. However, the mass percent compositions of compounds also affect the MS isotopic pattern, which can be readily resolved by conventional HMRP instruments to provide a wealth of class-specific information.

Another important intrinsic property of molecules that has often been utilized in mass spectrometry, for both small [35] and large molecule [61] assignments, is the molecular isotopic pattern. Heavy isotopes of a given element uniquely contribute to heavy isotopologue peak intensities as molecular masses of analytes increase. For example, the second most abundant isotopes of carbon and nitrogen are ^{13}C (1.1% natural abundance (NA)) and ^{15}N (0.36% NA), respectively, which are ~1 Da heavier than their corresponding most abundant counterparts (^{12}C and ^{14}N). Hence, presence of either ^{13}C or ^{15}N increases the molecular mass of an ion by one atomic mass unit (*i.e.*, designated as A+1 elements that contribute to A+1 peak in a mass spectrum). Because of the larger NA contribution from ^{13}C (i.e., 1.1%) than ^{15}N (i.e., 0.36%) as well as greater mass proficiency (0.003355 for ^{13}C and 0.000109 for ^{15}N), contributions to average molecular weights are larger from ^{13}C isotopes than ^{15}N isotopes; moreover, biomolecules generally contain larger numbers of carbon atoms than nitrogen atoms and

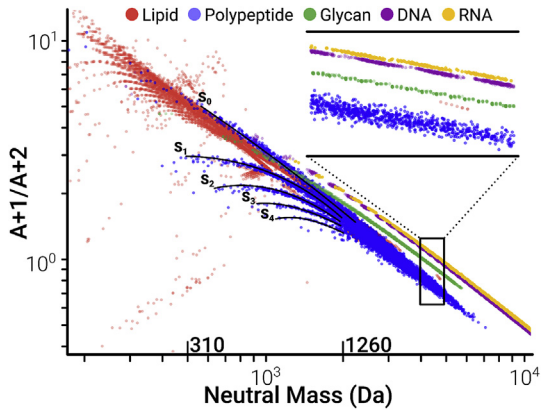


Fig. 2. Log-log plots for theoretical MS isotopologue peak ratio, A+1/A+2, as a function of compound neutral mass (Da), show high degrees of separation between classes of biomolecules. Predominantly linear trends for each class show general correlations between the isotope ratios and neutral masses. Polypeptide sulfur content is also distinguished between the 310–1260 Da mass range as A+1/A+2 decreases with each sulfur atom inclusion (shown by blue trendlines labeled S_0–S_4). For the sake of figure clarity, the lowest and highest mass compounds are omitted. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

hence A+1 peaks (mostly contributions from ^{13}C) are often informally referred to as ^{13}C isotope peaks. It should be noted that, when present, several other isotopes can also contribute to A+1 peak (*e.g.*, ^{2}H, ^{17}O, *etc.*); UHMRP required to resolve these "fine structures" within the observed isotopic envelopes is beyond the reach of conventional mass spectrometers [36,62,63]. Therefore, observed A+1 peaks in mass spectra, that might be from a combination of various isotopes, are often unresolved. Similar to A+1 elements (*e.g.*, ^{13}C, ^{14}N, and ^{2}H), the second most abundant isotopes contribute to the observed relative abundance of A+2 peaks. For instance, oxygen and sulfur (*i.e.*, ^{18}O (0.21% NA) and ^{34}S (4.3% NA)) are ~2 Da heavier (*i.e.*, A+2 elements) than their corresponding most abundant counterparts (^{16}O and ^{32}S) and thus contribute to A+2 peak. It should be noted that there are several other isotopes and numerous other combinations (*e.g.*, $^{2}H_2$, $^{13}C_2$, $^{15}N_2$, $^{2}H_1 + {}^{13}C_1$, $^{2}H_1 + {}^{15}N_1$, $^{13}C_1 + {}^{15}N_1$, *etc.*) that can also contribute to A+2 and other higher mass isotopologues (*e.g.*, A + 3, A + 4, …., A + n, where n is the number of observed peaks within an isotopic envelope for an analyte). Therefore, the relative contributions of each element to the mass of a compound affects isotopologue peak relative intensities in the mass spectrum. Fig. 2 displays logarithmic plots of (A+1/A+2) peaks for nucleic acid, glycan, polypeptide, and lipid biomolecules as a function of their neutral monoisotopic molecular (or A) masses. Given the relative contributions of different heavy isotopes to A+1 and A+2, the ratio A+1/A+2 indirectly relates the quantities of carbon, nitrogen, hydrogen, and other A+1 elements to oxygen, sulfur, and other A+2 elements.

The isotopic ratio plot (*i.e.*, A+1/A+2 as a function of neutral mass) presented in Fig. 2 provides one example (of many possible ways) of how such visual displays can be used for biomolecule class differentiation; multidimensional raw data used to train FFNNs allow access to numerous other A+1/A+2 type plots (ratios of other isotopic peaks) and "hidden" relationships that are not easily discernable by using two-dimensional plots. For this discussion, the logarithmic plot of A+1/A+2 as a function of compounds' neutral mass (Fig. 2) is helpful in so far as A+1/A+2 is influenced by both A+1 and A+2 elements. As expected, amplitude of the A+1/A+2 ratio decreases with increasing neutral mass up to certain values for different compounds — or a chemical class-dependent value. The initial decrease is due to the generally increased presence of A+2 elements (*i.e.*, oxygen and sulfur) and the probability of larger molecules containing multiple A+1 element heavy isotopes (*e.g.*, $^{2}H_2$, $^{13}C_2$, $^{15}N_2$, $^{2}H_1 + {}^{13}C_1$, $^{2}H_1 + {}^{15}N_1$, $^{13}C_1 + {}^{15}N_1$, *etc.*). Observed biomolecule class-dependent trends vary as a function of elemental mass composition. The top right inset in Fig. 2 shows the expanded region from ~2000 to 2500 Da; although the parallel trends for different compound classes are close to each other, they are fully separated and correlate to the differences in mass percent compositions observed in Fig. 1 that can be used for class differentiation. For instance, the ribonucleic acid (RNA, yellow) trendline is slightly higher than deoxyribonucleic acid (DNA, purple) because of the loss of CH_2 from ~25% of RNA residues (thymine exchanged with uracil). However, the magnitude of the RNA shift is reduced by the gain of oxygen at the ribose 2' position. From ~310 to 1260 Da (indicated number labels on the x-axis of Fig. 2), the sulfur content of the polypeptides (Fig. 2, blue) can be visually distinguished by the divergent trendlines (labeled S0–S4) due to the major contribution of ^{34}S (4.21% relative abundance) to A+2 [34]. The polypeptides along the S_0 trend contain no sulfur, and each descending trend (S_1–S_4) incorporates collections of polypeptides that contain an additional sulfur atom. Also, it should be noted that because of the polypeptide diversity (polymer growth possibilities to yield similar MW as compared to limited polymer growth possibilities for glycan, DNA, or RNA classes) and possibility of having various elemental compositions yielding similar polypeptide masses, the blue trendline in the expanded region of Fig. 2 is wider (in the y-axis, corresponding to (A+1/A+2) values) than the counterpart trendlines for glycans, DNAs, and RNAs. Such inter- and intra-class variations of isotopic patterns might be difficult to discern without the use of neural networks that often capitalize on "hidden" relationships for class discriminations.

The applicability of isotopic envelope features in their use by FFNNs can be limited by both instrumental and analyte-specific constraints. In the study presented here, the isotopic envelope was centroided and each peak was integrated to reduce the complexity of the FFNN input layer and eliminate the effects of variance of resolution across MS instruments. The centroided monoisotopic *m/z* value (input 1) was chosen for the high correlation of *m/z* with isotopic ratios in organic compounds (Fig. 1). However, with respect to macromolecules such as proteins that contain a large number of carbons, the monoisotopologue peak ($^{12}C_{all}$) relative intensity (input 2) is often very low (*i.e.*, the probability of having a large protein molecule with all of its carbons as ^{12}C) and often below the instrument detection limits. In our approach, the proper classification of biomolecules with FFNNs is dependent on successful identification and measurements of the monoisotopologue MS peak. Thus, application of the current approach to experimental data is limited to species with observable monoisotopologue peaks. Likewise, the measurement of the isotopologue peak relative intensity inputs, A (input 2), A+1 (input 3), and A+2 (input 4), and thus the calculated isotopic ratios, A/A+1 (input 5) and A+1/A+2 (input 6), are subject to the same limitations of instrumental sensitivity and MRP. It could be argued that the calculated isotopic ratios (inputs 5 and 6) are unnecessary since the FFNN training could "discover" those relationships; however, the inclusion of these inputs improved FFNN training performance in all cases. As the fourth (and onwards) isotopologue peak(s) are often below instrument detection limits for low mass (and low abundance) analytes in ESI-MS experiments, these peak intensities were not included as inputs in our specific FFNN training sets. However, depending on the molecular weight (or *m/z*) ranges of interest, it is possible to utilize signals from other isotopologues (e.g., other relatively high abundance species) as inputs. In other words, two important criteria in considering a particular set of isotopologues to select for FFNN training sets are (a) relative

abundance consideration (signal-to-noise consideration for MS detectability) and (b) computational cost (e.g., additional input may not necessarily improve the success rate sufficiently large but require unreasonably large computational times). The selection of three isotopologues in the presented study provided a good balance between the desired success rate and computational cost. The calculation of KMD (input 7) is also contingent on the detection and mass measurement accuracy of the monoisotopic *m/z*. The KMD input (7) was included for its ability to assist in MS ion classification challenges and for its contribution to network training performance improvements [55]. Within the context of small molecule analysis (*e.g.*, metabolomics, lipidomics, and peptidomics), the above features are readily distinguished in most modern TOF-MS workflows.

3.2. FFNN training and performance

For PEPNET, the "non-polypeptide" portion of the supervised training set consisted of lipid, glycan, polar metabolite, and nucleic acid input vectors (i.e., one-dimensional data arrays containing network input values; N_{Total} = 23,816 post-growth method; N_{Train} = 16,673 (70%); N_{Test} = 3543 (15%); $N_{Validation}$ = 3600 (15%)). The "polypeptide" portion of the supervised training set consisted of polypeptide input vectors (N_{Total} = 23,816 post-growth method; N_{Train} = 16,669 (70%); N_{Test} = 3602 (15%); $N_{Validation}$ = 3545 (15%)). PEPNET had a consistent mean percent error of 7.2% for each training, validation, and test datasets; in other words, PEPNET incorrectly classified 7.2% of inputs submitted in the training set. For the supervised training test set, T_P and T_{NP} had true positive rates (TPR, percentage of actual positives classified as such) of 94.2% and 91.4%, respectively, and positive predictive values (PPV, percentage of actual positives in all predicted positives) of 91.7% and 93.9%, respectively. For LIPNET, the "non-lipid" portion of the supervised training set consisted of polypeptide, glycan, polar metabolite, and nucleic acid input vectors (N_{Total} = 23,809 post-growth method; N_{Train} = 16,712 (70%); N_{Test} = 3552 (15%); $N_{Validation}$ = 3545 (15%)). The "lipid" portion of the supervised training set consisted of lipid input vectors (N_{Total} = 23,809 post-growth method; N_{Train} = 16,620 (70%); N_{Test} = 3591 (15%); $N_{Validation}$ = 3598 (15%)). The LIPNET training, validation, and test datasets had 12.1%, 12.6%, and 12.1% overall percent errors, respectively. T_L and T_{NL} had TPRs of 92.1% and 83.7%, respectively, and PPVs of 85.1% and 91.3%, respectively, for the supervised training test set. The training and validation confusion matrices [64] for PEPNET and LIPNET varied by $\leq$ 1%, suggesting that the networks were not overfit (for more information regarding PEPNET and LIPNET training performance metrics, see Appendix B.4 and B.5). PEPNET performed slightly better than LIPNET presumably due to the narrower variety of possible peptide chemical compositions as a function of mass relative to lipids, which exhibit great structural and compositional variety. LIPSUBNET was an 8-target classification network trained to classify previously-assigned T_L input vectors into T_{FA}, T_{GL}, T_{GP}, T_{PK}, T_{PR}, T_{SP}, T_{SL}, and T_{ST} target classes (*i.e.*, a comprehensive set of targets corresponding to all lipid subclasses). The LIPSUBNET training, validation, and test datasets had 7.9%, 8.1%, and 8.2% overall percent error, respectively. Respective TPR and PPV values for each target and predicted class is provided in a confusion matrix diagram in Appendix B.6.

Networks trained with multiple output nodes (e.g., LIPSUBNET) do not exhibit equal TPR and PPV values across each output node. Therefore, we utilized a parameter designated as "confidence threshold" to reduce variance in network performance characteristics (i.e., TPR and PPV) across network output nodes and improve the accuracy of the trained networks. Confidence threshold, a user-defined scalar value from 0.0 to 1.0, constrained "confident" (unassociated with confidence interval) identifications to those that result from network scores higher than the confidence threshold value. Classifications with network scores less than the user-defined confidence threshold are removed from consideration in calculation of network performance characteristics. For example, with a defined confidence threshold of 0.7, a classification that resulted from a score of less than 0.7 would be excluded; conversely, a classification of that resulted from a score of 0.7 or higher would be included in calculation of network performance characteristics. Fig. 3 displays the TPR of LIPSUBNET for all unique training set inputs as a function of confidence threshold. The TPR at a 0.0 confidence threshold is not uniform across class targets, ranging from 77.9% (prenol lipids, Fig. 3, trace number 8 in blue-green) to 98.6% (saccharolipids, Fig. 3, trace number 1 in light green). The low performance of prenol lipid classification can be attributed to relatively high chemical structural similarity to the likes of sterol lipids (Fig. 3, trace number 7 in orange), both of which share a common biosynthetic pathway [65], and, to a lesser extent, fatty acyls (Fig. 3, trace number 6 in red) that are likewise mostly comprised of minimally branched hydrocarbon chains [65]. As confidence threshold is increased, the TPR generally increased for each class as more incorrect classifications with lower network output scores were removed than correct classifications (which resulted from generally higher network output scores). As shown in Fig. 3, the lowest performing classes, prenol lipids (trace number 8 in turquoise), sterol lipids (trace number 7 in orange), and fatty acyls (trace number 6 in red), exhibit the largest improvements in TPR as a function of confidence threshold. By removing low scoring, incorrect predictions in these classes, the variance in performance between all classes is reduced. For example, at confidence threshold 0.80, all classes exhibit TPRs of greater than or equal to 95%.

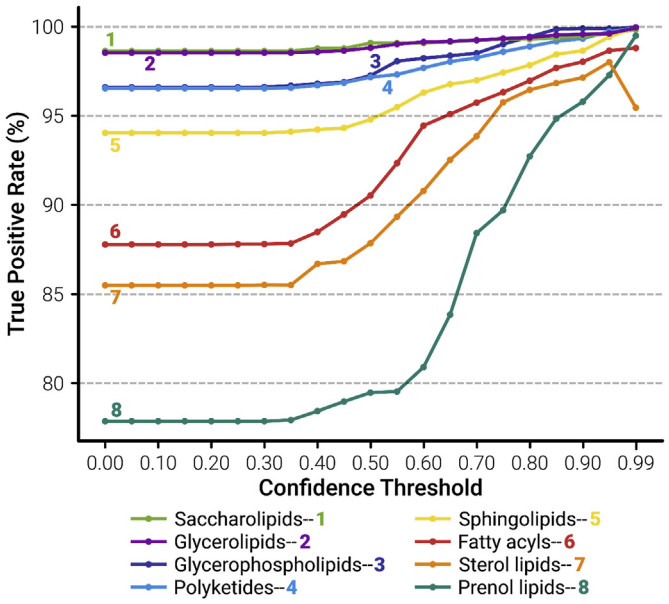


Fig. 3. True positive rates for 8 trained networks (each aimed at a specific chemical class identification) for the computationally generated data at confidence thresholds (0–1.0) are shown. The true positive rate (TPR (%)) is defined as the percentage of actual positives that were classified as such. Increasing the confidence threshold value causes TPR to approach 100% for most classes. Sterol lipids (7, orange) exhibit a decrease in TPR in a rare case where more correct than incorrect classifications are removed by an increase in confidence threshold (additional explanations are provided in the Results and Discussion section and Supplemental). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

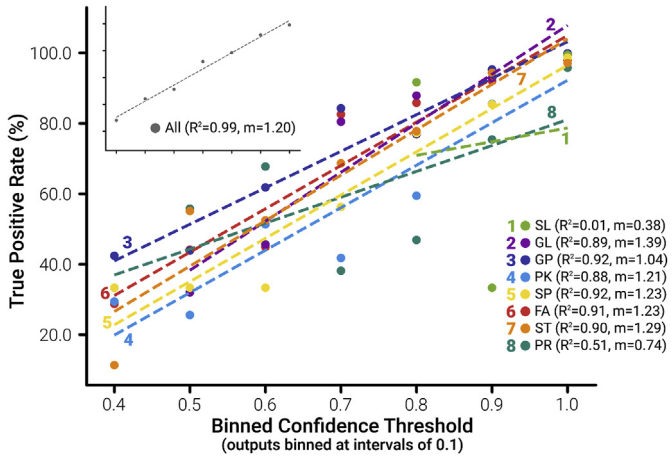
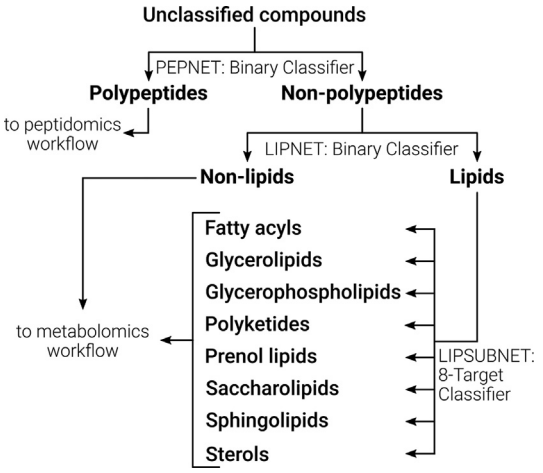


Fig. 4. Visualization of the true positive rate (TPR) of LIPSUBNET network outputs binned at confidence threshold intervals of 0.1 (e.g., (0.1, 0.2], (0.2,0.3],…(0.9, 1.0], where each plotted point represents the TPR of the outputs which scored in the binned interval range; only bin intervals that contained at least 10 outputs were plotted. Subclasses are represented in the legend as follows: saccharolipids (SL, 1), glycerolipids (GL,2), glycerophospholipids (GP, 3), polyketides (PK, 4), sphingolipids (SP, 5), fatty acyls (FA, 6), sterol lipids (ST, 7), and prenol lipids (PR, 8). The "All" category (inset, with the identical x and y axis ranges of 0.4–1.0 binned intervals and 0–100% TPR) represents the combined lipid classification accuracy by LIPSUBNET.



Scheme 1. Representative *in silico* fractionation neural decision tree workflow diagram of a biological sample dataset containing polypeptide, lipid, and polar metabolite components.

At high confidence thresholds (*i.e.*, >0.95), it is possible for TPR to decrease; this is shown by the sterol lipid class that dropped by 2.56% TPR when the confidence threshold was increased from 0.95 to 0.99 (Fig. 3, trace number 7 in orange). Decreases in TPR as a function of confidence threshold occur as more correct classifications than unconfident assignments are removed from the output. This behavior results from the networks' occasional proclivity to make confident, incorrect assignments primarily in the case of highly similar classes; for example, the sterol lipid class (orange, trace number 7 in Fig. 3) exhibited decreases in accuracy between confidence threshold 0.95 and 0.99 in 8/10 trained networks due to high scoring (>0.90), false positive classifications of prenol lipid inputs into the sterol lipid class.

Additionally, it is important to note that each class loses coverage (*i.e.*, the percentage of retained confident classifications of all classifications) as a function of confidence threshold as unconfident classifications are removed. As a rule, increased TPR as a consequence of higher confidence threshold selection results in some loss of coverage. The degree of coverage loss depends on the number of classifications removed; therefore, large increases in TPR result in large losses in coverage. The variance in class coverage is reported as estimated standard error of the mean (SEM % $= \frac{s}{\sqrt{n}} \times$ 100%). Based on the LIPSUBNET results, fatty acyls, sterol lipids, and prenol lipids lost coverage (and gained TPR; Fig. 3 traces 6, 7, and 8, respectively) at greater rates, causing high variance in coverage (76.6 ± 6.9%) between classes' TPR at a confidence threshold of 0.90. The relationships between class coverage and confidence threshold for each class in LIPSUBNET is displayed in Appendix B.7. The inverse relationship between TPR and coverage constitutes a "trade-off" that the user must manage depending on the user's specific needs for increased classification accuracy or higher coverage of all detected analytes. For general use of LIPSUBNET in this demonstration, the authors suggest use of a confidence threshold of 0.70 that balances TPR (96.3 ± 1.3%), total percent coverage (88.7 ± 3.9%), and uniformity of class representation.

The general positive trends for observed TPR values as a function of confidence threshold (after a confidence threshold of ~0.30; Fig. 3) suggests a relationship between the magnitude of network output score and the probability that an actual positive is classified as such. To visualize the relationship between network output score and TPR without the influence of high scoring outputs at every threshold (as in Fig. 3), the TPR of network output scores binned at confidence threshold intervals of 0.1 (*e.g.*, [0.0, 0.1], (0.1, 0.2], …, (0.9, 1.0] bins for the 0.4 to 1.0 confidence threshold range) are displayed in Fig. 4 (where only sufficiently populated bins with $n \geq 10$ inputs have been included). The raw network outputs of all lipid subclasses (x values in Fig. 4) exhibit an appreciable linear correlation (positive m values from 0.38 to 1.39, Fig. 4) with TPR; the mean R^2 value for all classes (excluding saccharolipids, SL with $R^2 = 0.01$, trace number 1 of Fig. 4 in light green) is 0.85 ± 0.06 (±SEM). The exceptionally low R^2 of 0.01 and high TPR (of 98% as seen in Fig. 3, trace 1 in light green) for SL indicate that LIPSUBNET may have been overfit for characterization of the SL target class. However, when all (binned) lipid outputs were considered together, a high linear correlation between accuracy and network output score was observed ($R^2 = 0.99$, Fig. 4 inset). The slope (m) of the combined linear regression trend line (inset, with the identical x and y axis ranges, in Fig. 4) was 1.20, suggesting a near 1:1 relationship between the magnitude of output score and the probability that a classification is correct; as such, the magnitude of raw network outputs may be useful for determining classification confidence levels.

3.3. *In silico fractionation of multi-class component mixtures*

The supervised training sets of PEPNET, LIPNET, and LIPSUBNET were constructed to demonstrate the *in silico* fractionation (iSF) approach. In this case, we present an application of the iSF approach that utilizes a NDT structure to parse a sample dataset containing lipid, polypeptide, and polar metabolite components (Scheme 1). A visual representation of the iSF approach to an artificially combined, representative multi-class component experimental LC-MS data is presented in Fig. 5. The analyzed LC-MS dataset (represented in Fig. 5) is not intended to reflect realistic sample preparation or chromatographic conditions as all analytes were initially prepared, separated, and detected in ideal conditions. However, it was prepared to demonstrate a use case for iSF in which analyte signal could not be assumed to originate from a single class. It should also be noted that, with exception to enhancements to detection sensitivity, sample preparation and chromatographic conditions do not uniquely affect the spectral features of the MS isotopic envelope utilized by iSF. Additionally, as Scheme 1 and

Fig. 5 suggest, the end goal of this work is to guide fractionation of the full MS dataset in state amenable for separate downstream analyses for classes of molecules in a multi-class mixture that have unique data processing requirements. However, the challenges associated with full integration of iSF into a multi-omics workflow will be addressed in a future study.

Raw network output scores for PEPNET, LIPNET, and LIPSUBNET related to Fig. 5 are provided in Appendix Table C1. Fig. 5 demonstrates the application of iSF to the classification of multiple types of eluting compounds in a standard LC-ESI-MS analysis using experimentally gathered data. Each eluting compound is pictorially represented by its detected LC peak profile. Peaks that have yet to be positively classified (i.e., classified into a discrete category of biomolecule) were shown in gray and were assigned a color once positively classified. The iSF approach begins with generating an input vector for each of the unclassified compounds in a LC-MS dataset (Fig. 5, top, gray). As the first step in Fig. 5, PEPNET proceeds to classify network inputs (second row, Fig. 5) as polypeptides (T_P, green) or non-polypeptides (T_{NP}, gray). In the dataset presented in Fig. 5 (*i.e.*, the combined dataset discussed at the end of the Experimental section comprised of an eluting mixture of brain peptides, lipids, and polar metabolites detected by ESI-MS), polypeptides were classified by PEPNET with 100% TPR. For the second step (row 3, Fig. 5) LIPNET classified the input vectors that PEPNET classified as T_{NP} (gray peaks in row 2 of Fig. 5) as either lipids (T_L, red) or non-lipids (T_{NL}, gray) as shown in row 3 of Fig. 5 with 100% TPR. The vector inputs classified as T_L by LIPNET were then submitted to LIPSUBNET for classification into lipid subclasses (*i.e.*, subclasses listed by LipidMaps). LIPSUBNET classified 37 out of 39 inputs correctly (95% TPR for all class targets). See Table 1 for statistical metrics for each class; FA, GL, GPL, PK, PR, SL, SP, and ST acronyms in Table 1 (column headers) are used for fatty acyl, glycerolipid, glycerophospholipid, polyketide, prenol lipid, saccharolipid, sphingolipid, and sterol lipid, respectively, where row headers TPR, TNR, FPR, and FNR stand for true positive rate, true negative rate, false positive rate, and false negative rate, respectively. One sterol lipid input vector was incorrectly classified as a glycerolipid with a network score of 0.597 and one prenol lipid input vector was incorrectly classified as a fatty acyl with a network score of 0.622. However, for both the sterol and prenol misclassifications, the true class targets received the second highest network scores of 0.323 and 0.217, respectively. If the previously suggested confidence threshold (of 0.70) is applied to the results of the iSF workflow reported herein, 5 classifications, including both incorrect classifications, are removed, resulting in 100% TPR with 87% (*i.e.*, 34 out of 39 input) coverage. Additionally, input vectors previously classified by PEPNET or LIPNET may be resubmitted to the other (e.g., input vector classified as T_P by PEPNET can be reclassified by LIPNET) for secondary confirmation. All positively classified input vectors resubmitted to either PEPNET or LIPNET (respective to their previous classification) were classified with 100% TPR.

To demonstrate an example of iSF's application to a real-world multi-class separation, PEPNET and LIPNET were used to analyze a separated mixture of trypsin digest peptides and lipids. The total ion LC chromatogram of the combined lipid and peptide separation is shown in Fig. 6. The dotted lines mark the LC retention time boundaries between which, from left to right, peptides, unknowns,

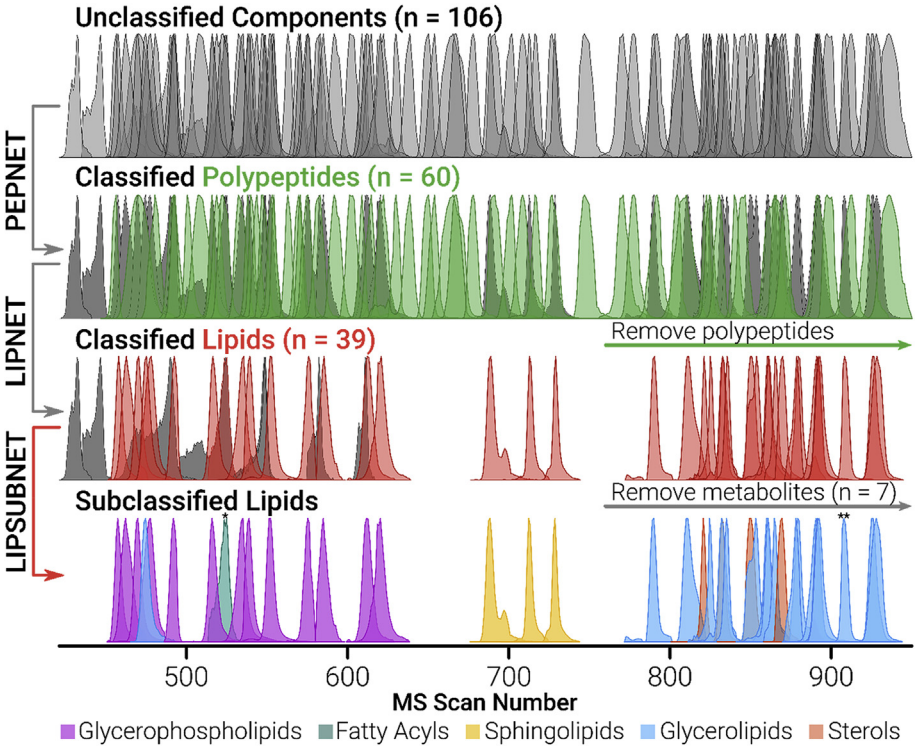


Fig. 5. A visual representation of the *in silico* fractionation approach to a randomly selected portion of an LC-MS dataset that included a total of 106 unknown analytes. The original input data (106 Gy LC peaks) in top row was created by adding a portion of an experimentally acquired LC-MS output from a rat brain tryptic digest analysis to lipidomics LC-MS data downloaded from Chorus) [43]. Unclassified components (gray LC peaks) were submitted to PEPNET which classified inputs as 60 polypeptides (T_P, LC peaks in second row from the top) or 46 non-polypeptides (T_{NP}, gray LC peaks). T_P inputs were removed, and the remaining T_{NP} inputs (46 species) were submitted to LIPNET for a second chemical class identification, which classified inputs as 39 lipids (T_L, red LC peaks in third row from the top) and 7 non-lipids (T_{NL}, gray LC peaks in the third row from the top). T_{NL} inputs were removed, and T_L inputs were submitted to LIPSUBNET, which classified T_L inputs into narrower lipid subclasses. PEPNET, LIPNET, and LIPSUBNET had true positive rates of 100%, 100%, and 95%, respectively. The * and ** symbols represent a prenol lipid component and a sterol component that were erroneously predicted as fatty acyl and glycerolipid, respectively (additional details on mis-classified components are provided in the Results and Discussion section). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
LIPSUBNET statistical metrics for classification of experimental data.

Metric	FA	GL	GPL	PK	PR	SL	SP	ST
TPR	–	94.7%	100.0%	–	0.0%	–	100.0%	80.0%
TNR	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
FPR	2.6%	4.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
FNR	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	20.0%
Overall	Accuracy	98.7%		TPR	94.9%		TNR	99.3%

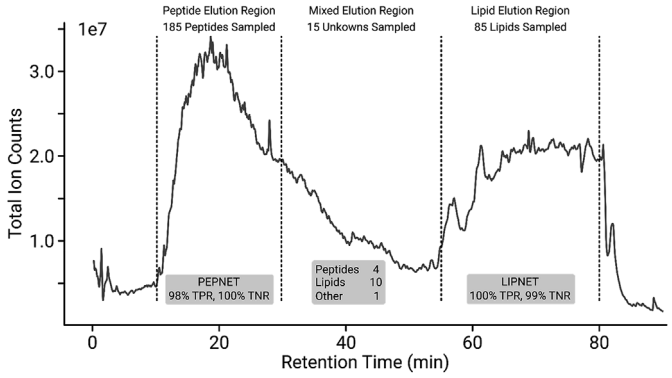


Fig. 6. A visual representation of the *in silico* fractionation approach to a portion of a reversed-phase LC-MS dataset in which a mixture of HeLa digest peptides (total of 150 ng) and rat brain-extract lipids (from 250 μg rat brain) were separated and analyzed in the same acquisition. Known peptide data were extracted from between 10 and 30 min, lipid data from between 55 and 80 min, and unknowns from the middle region where both hydrophobic peptides and hydrophilic lipids were expected to elute. Regarding the classification of known components, PEPNET had a 98% TPR and 100% TNR, and LIPNET had a 100% TPR and 99% TNR. In the middle region, 4 unknowns were classified as peptides, 10 were classified as lipids, and 1 was classified as a non-lipid and non-peptide. No conflicting classifications were made for any unknown component. The period before 10 min was excluded as it contained only unretained components and background signal. The period after 80 min was excluded as it contained only a strong contaminate/polymer signal.

and lipids eluted. In this demonstration, the true classes of each of analyte signals were determined by their respective reverse-phase chromatographic elution times.

As indicated in Fig. 6, LC eluting analytes between 10 and 30 min were determined to be peptides, and those that eluted between 55 and 80 min were determined to be lipids. The classification of each analyte was confirmed by independent LC-MS analyses of the peptide and lipid mixtures. Polypeptide analytes were classified by PEPNET as T_P with 98% TPR (181 out of 185 inputs). Of the 185 polypeptide components, 4 polypeptides were incorrectly classified by PEPNET as T_{NP}. Each of the 4 polypeptides were small (MW < 1000 Da), singly-charged ions; however, it should be noted that most other small, singly-charged polypeptide components were classified correctly (12 out of 16 inputs). Lipid inputs were classified by LIPNET as T_L with 100% TPR (85 out of 85 inputs). Also, these lipid inputs submitted to PEPNET were classified as T_{NP} with 100% TNR. Peptide inputs submitted to LIPNET were classified as T_{NL} with 99% TNR (183 out of 185 inputs). The two polypeptide inputs that were misclassified by LIPNET as T_L were 2 of the 4 small, singly-charged polypeptide inputs misclassified by PEPNET. In the mixed elution region from 30 to 55 min, 15 unknown components were classified by both PEPNET and LIPNET. In this group of inputs, iSF classifications were self-consistent as 4 inputs were classified as T_P and T_{NL} (designated as "T_P/T_{NL}"), 10 inputs as T_L/T_{NP}, and 1 input that was classified as neither a peptide nor lipid (both T_{NP} and T_{NL} or T_{NP}/T_{NL}). No conflicting or self-inconsistent classifications (i.e., belonging to both T_P and TL classifications) were made for any unknown input. Raw network output scores for PEPNET and LIPNET related to Fig. 6 are provided in Appendix Table C.2.

3.4. Considerations for future applications

The generalization capabilities of FFNNs enabled PEPNET, LIPNET, and LIPSUBNET to accommodate non-trivial variance in the *m/z* domain. The supervised training sets for each class were constructed with elemental compositions corresponding to the commonly observed adduct ions in ESI (*e.g.*, $[M+H]^+$ for most polypeptides, $[M+Na]^+$ for glycans, $[M-H]^-$ for oligonucleotides, *etc.*). However, there were a variety of detected lipid adduct forms (*e.g.*, $[M+NH_4]^+$, $[M+ACN+H]^+$ $[M+Na]^+$, *etc.*), that were successfully classified by LIPSUBNET and LIPNET, even though the training sets of lipids for LIPSUBNET and LIPNET were restricted to protonated adduct lipid forms. Given the easy-to-obtain nature of the selected mass spectral vector inputs (*i.e.*, isotopologue peak relative intensity, exact mass, and KMD information) and the insensitivity of iSF to potential interferences from adduct ions, iSF could be applied to most broadband MS workflows in which analytes' molecular ion isotopic envelopes are preserved. Given its robustness, iSF can be applied as a pre-processing step in conceivably any concurrent multi-omics analysis (in which MS1 spectra are acquired) to provide crucial information about analyte classes and guide future data processing. Such capabilities should be useful for characterization of complex biological systems such as bacterial differentiation in microbiome studies [66], class identification in biological MS imaging, and pictorial representation of biomarker panels (healthy controls vs disease states) in clinical studies. Although the presented work here has focused on positive-ion mode experiments and classification of intact molecular species, iSF can be applied to other complementary types of data. For example, in future contributions, we plan to evaluate the performance of iSF for utilizing data acquired under negative-ion mode and classification of fragments and other modified structures (such as metal adducts).

The neural network training period took several seconds to several minutes (for details, see Appendix Figures B1 and B.2); however, once trained, each network classification was rapid (~85 μs of processor time). Thus, these operations can be performed at frequencies comparable to TOF data acquisition rates, which (depending on the measured *m/z* range) can range from ~10 to 100 kHz [48,67,68].

The presented approach is robust in dealing with instrumental noise and small variations in analytes' measured masses (data acquisition restrictions that are due to formation of various types of adducts, resolving power limits, and mass measurement errors). Additionally, our findings suggest that iSF can be applied successfully to MS data produced from low (L) MRP instruments. Versions of LIPNET, PEPNET, and LIPSUBNET were trained and tested with monoisotopic *m/z* and KMD values restricted to only two decimal places. Likewise, the monoisotopic *m/z* and KMD values of the experimental test inputs from the artificially combined multi-omics separation were restricted to two decimal places. These two types of networks (viz., networks trained using data sets with either (a) only two or (b) four decimal places for *m/z* values) had nearly identical performance characteristics both in application to computationally generated data and experimental MS data to LIPNET, PEPNET, and LIPSUBNET. The exciting implication of this

finding is iSF's applicability to data from relatively LMRP instruments (with precision to only 2 decimal places) that exhibit mass measurement errors unsuitable for elemental composition determination (e.g., data with mass measurement errors larger than 10 ppm). Tabulated performance metrics for the networks trained with precision to 2 decimal places are shown in Appendix Table C.3.

Because relative peak intensities and isotopic ratios are the primary dimensions of separation used by the neural network, the proposed approach is sensitive to convolution of analyte signals in the *m/z* domain (hence, peak capacity limits in the *m/z* dimension govern the level of sample complexity that can be tolerated). We recommend the use of pre-MS physical separations such as LC and/or IM to prevent potential peak convolutions in the *m/z* domain and increase isotopic envelope purity. IM profiles or trendlines for biomolecules have also shown class-dependent trends that may assist in classifications [32]; however, overlaps of *m/z*-mobility trends complicate classification of some groups of biomolecules [32]. It should be noted that, even with physical separation prior to MS injection, convolution of closely related chemicals across measurement domains is possible. In such instances of partial isotopic envelope convolutions of closely related lipids in the LC and *m/z* domains (for example, as reported in Fig. 6) iSF is still able to correctly classify each convolved component. For instance, 11 of the 85 lipid inputs (all of which were correctly classified as T_L by LIPNET) were partially convoluted in the LC retention time and *m/z* domains. Of the 11 convolved lipids, 6 were partially convolved by closely related species that differed by a degree of unsaturation. In regions of full LC convolution (e.g. retention times at which two lipids that differ by one degree of unsaturation are eluting; an example of which is shown in Appendix B.8), the third isotopologue peak of the singly-charged, unsaturated species (which eluted first in each case) was fully convolved with the monoisotopologue peak of the saturated species (which elutes second in each case). However, by sampling from the leading edge of the saturated species' LC peak and the trailing edge of unsaturated species' LC peak, sufficiently pure isotopic envelopes were obtained. The other 5 convolved lipid components were partially convolved in the *m/z* domain (i.e., each isotopologue peak convolved up to ~20% peak height; an example for which is shown in Appendix B.9), but extracted isotopic features were sufficiently pure for iSF to classify each component correctly.

4. Conclusions

In this study, we describe an *in silico* fractionation approach for classification of small biological compounds from MS data *via* isotopic ratio analysis using a neural decision tree. The FFNNs estimate the relationships that define and separate biomolecular classes (*e.g.*, lipids, glycans, polypeptides, *etc.*) based on their respective isotopic distribution patterns and, therefore, their elemental compositions. The FFNNs utilized to demonstrate iSF (PEPNET, LIPNET, and LIPSUBNET) were sensitive in their application to experimentally detected chemical components: PEPNET had a TPR of 100%, LIPNET had a TPR of 100%, and LIPSUBNET had a combined TPR of 95% (without confidence threshold constraint). The specific demonstration presented here, constitutes one possible design and application of iSF; however, depending on the sample composition and class types present in the sample, the iSF workflow can be tailored for a wide variety of multi-class component mixtures.

Funding

This work was supported by the National Science Foundation [grant numbers: IDBR-1455668 and CHE-1709526].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Luke T. Richardson: Methodology, Investigation, Software, Formal analysis, Data curation, Validation, Visualization, Writing - original draft. **Matthew R. Brantley:** Software, Formal analysis, Visualization, Writing - review & editing. **Touradj Solouki:** Funding acquisition, Conceptualization, Project administration, Methodology, Supervision, Formal analysis, Visualization, Writing - review & editing.

Acknowledgements

The original concept for iSF was developed by TS as a homework assignment for CHE 5V60 Course (Special Topics in Advanced Analytical Chemistry). The authors acknowledge Ian Anthony (Baylor University) for his assistance with code refactoring and manuscript review, and Dr. Christopher M. Kearney, Dr. S. M. Ashiqul Islam, and Reese Martin for helpful discussions on network selections.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.aca.2020.02.036.

References

[1] G. Sarah, J.A. Reisz, N. Travis, K.C. Hansen, D.A. Angelo, Characterization of rapid extraction protocols for high-throughput metabolomics, Rapid Commun. Mass Spectrom. 31 (2017) 1445–1452.
[2] N. Travis, K.C. Hansen, D.A. Angelo, A three-minute method for high-throughput quantitative metabolomics and quantitative tracing experiments of central carbon and nitrogen pathways, Rapid Commun. Mass Spectrom. 31 (2017) 663–673.
[3] J.C. Guder, T. Schramm, T. Sander, H. Link, Time-optimized isotope ratio LC–MS/MS for high-throughput quantification of primary metabolites, Anal. Chem. 89 (2017) 1624–1631.
[4] A. Lubin, S. Geerinckx, S. Bajic, D. Cabooter, P. Augustijns, F. Cuyckens, R.J. Vreeken, Enhanced performance for the analysis of prostaglandins and thromboxanes by liquid chromatography-tandem mass spectrometry using a new atmospheric pressure ionization source, J. Chromatogr. A 1440 (2016) 260–265.
[5] A. Lubin, R. De Vries, D. Cabooter, P. Augustijns, F. Cuyckens, An atmospheric pressure ionization source using a high voltage target compared to electrospray ionization for the LC/MS analysis of pharmaceutical compounds, J. Pharmaceut. Biomed. Anal. 142 (2017) 225–231.
[6] W. Ai, H. Nie, S. Song, X. Liu, Y. Bai, H. Liu, A versatile integrated ambient ionization source platform, J. Am. Soc. Mass Spectrom. 29 (2018) 1408–1415.
[7] W. Zhang, F. Li, L. Nie, Integrating multiple 'omics' analysis for microbial biology: application and methodologies, Microbiology 156 (2010) 287–301.
[8] S.-Y. Ahn, N. Jamshidi, M.L. Mo, W. Wu, S.A. Eraly, A. Dnyanmote, K.T. Bush, T.F. Gallegos, D.H. Sweet, B. Palsson, S.K. Nigam, Linkage of organic anion transporter-1 to metabolic pathways through integrated omics-driven network and functional analysis, J. Biol. Chem. 286 (2011) 31522–31531.
[9] Y. Gong, R. Cao, G. Ding, S. Hong, W. Zhou, W. Lu, M. Damle, B. Fang, C.C. Wang, J. Qian, N. Lie, C. Lanzillotta, J.D. Rabinowitz, Z. Sun, Integrated omics approaches to characterize a nuclear receptor corepressor-associated histone deacetylase in mouse skeletal muscle, Mol. Cell. Endocrinol. 471 (2018) 22–32.
[10] J.E. Szulejko, Z. Luo, T. Solouki, Simultaneous determination of analyte concentrations, gas-phase basicities, and proton transfer kinetics using gas chromatography/Fourier transform ion cyclotron resonance mass spectrometry (GC/FT-ICR MS), Int. J. Mass Spectrom. 257 (2006) 16–26.
[11] B.M. Ruddy, C.L. Hendrickson, R.P. Rodgers, A.G. Marshall, Positive ion electrospray ionization suppression in petroleum and complex mixtures, Energy Fuel. 32 (2018) 2901–2907.
[12] S.L. Cohen, B.T. Chait, Influence of matrix solution conditions on the MALDI-MS analysis of peptides and proteins, Anal. Chem. 68 (1996) 31–37.
[13] B. Fuchs, J. Schiller, Recent developments of useful MALDI matrices for the

mass spectrometric characterization of apolar compounds, Curr. Org. Chem. 13 (2009) 1664–1681.
[14] M.E. Gimon, L.M. Preston, T. Solouki, M.A. White, D.H. Russell, Are proton transfer reactions of excited states involved in UV laser desorption ionization? Org. Mass Spectrom. 27 (1992) 827–830.
[15] R.T. Steven, J. Bunch, Repeat MALDI MS imaging of a single tissue section using multiple matrices and tissue washes, Anal. Bioanal. Chem. 405 (2013) 4719–4728.
[16] A.J. Schwartz, K.L. Williams, G.M. Hieftje, J.T. Shelley, Atmospheric-pressure solution-cathode glow discharge: a versatile ion source for atomic and molecular mass spectrometry, Anal. Chim. Acta 950 (2017) 119–128.
[17] J.H. Gross, Direct analysis in real time—a critical review on DART-MS, Anal. Bioanal. Chem. 406 (2014) 63–80.
[18] X. Dong, J. Cheng, J. Li, Y. Wang, Graphene as a novel matrix for the analysis of small molecules by MALDI-TOF MS, Anal. Chem. 82 (2010) 6208–6214.
[19] S.-C. Cheng, S.-S. Jhang, M.-Z. Huang, J. Shiea, Simultaneous detection of polar and nonpolar compounds by ambient mass spectrometry with a dual electrospray and atmospheric pressure chemical ionization source, Anal. Chem. 87 (2015) 1743–1748.
[20] L.M. Lang, P.W. Dalsgaard, K. Linnet, Quantitative analysis of cortisol and 6β-hydroxycortisol in urine by fully automated SPE and ultra-performance LC coupled with electrospray and atmospheric pressure chemical ionization (ESCi)-TOF-MS, J. Separ. Sci. 36 (2013) 246–251.
[21] L. Nyadong, A.S. Galhena, F.M. Fernández, Desorption electrospray/metastable-induced ionization: a flexible multimode ambient ion generation technique, Anal. Chem. 81 (2009) 7788–7794.
[22] A. Vaikkinen, B. Shrestha, J. Nazarian, R. Kostiainen, A. Vertes, T.J. Kauppila, Simultaneous detection of nonpolar and polar compounds by heat-assisted laser ablation electrospray ionization mass spectrometry, Anal. Chem. 85 (2013) 177–184.
[23] M.F. Mirabelli, R. Zenobi, Solid-phase microextraction coupled to capillary atmospheric pressure photoionization-mass spectrometry for direct analysis of polar and nonpolar compounds, Anal. Chem. 90 (2018) 5015–5022.
[24] C. Chen, L. Weng, K. Chen, F. Sheu, C. Lin, Symmetric atmospheric plasma source integrated with electrospray ionization for ambient mass spectrometry detections, IEEE Trans. Plasma Sci. 47 (2019) 1114–1120.
[25] W. Nie, L. Yan, Y.H. Lee, C. Guha, I.J. Kurland, H. Lu, Advanced mass spectrometry-based multi-omics technologies for exploring the pathogenesis of hepatocellular carcinoma, Mass Spectrom. Rev. 35 (2016) 331–349.
[26] C. Bock, M. Farlik, N.C. Sheffield, Multi-omics of single cells: strategies and applications, Trends Biotechnol. 34 (2016) 605–608.
[27] N. Ishii, M. Tomita, Multi-omics data-driven systems biology of E. coli, in: S.Y. Lee (Ed.), Systems Biology and Biotechnology of Escherichia coli, Springer Netherlands, Dordrecht, 2009, pp. 41–57.
[28] Y. Hasin, M. Seldin, A. Lusis, Multi-omics approaches to disease, Genome Biol. 18 (2017) 83.
[29] G. Siddiqui, A. Srivastava, A.S. Russell, D.J. Creek, Multi-omics based identification of specific biochemical changes associated with PfKelch13-mutant artemisinin-resistant plasmodium falciparum, J. Infect. Dis. 215 (2017) 1435–1444.
[30] K. Kaplan, S. Jackson, P. Dwivedi, W.S. Davidson, Q. Yang, P. Tso, W. Siems, A. Woods, H.H. Hill, Monitoring dynamic changes in lymph metabolome of fasting and fed rats by matrix-assisted laser desorption/ionization-ion mobility mass spectrometry (MALDI-IMMS), International Journal for Ion Mobility Spectrometry 16 (2013) 177–184.
[31] A.S. Woods, M. Ugarov, T. Egan, J. Koomen, K.J. Gillig, K. Fuhrer, M. Gonin, J.A. Schultz, Lipid/peptide/nucleotide separation with MALDI-ion mobility-TOF MS, Anal. Chem. 76 (2004) 2187–2195.
[32] L.S. Fenn, M. Kliman, A. Mahsut, S.R. Zhao, J.A. McLean, Characterizing ion mobility-mass spectrometry conformation space for the analysis of complex biological samples, Anal. Bioanal. Chem. 394 (2009) 235–244.
[33] A. Quach, B. Lomenick, A.J. Yoon, W. Cohn, J.P. Whitelegge, K.F. Faull, Omni-MS: a Method for Concurrent LC-MS Analysis of Electrolytes, Small Molecules, Lipids, Proteins, Nucleic Acids, and Polysaccharides, American Society for Mass Spectrometry Conference, Atlanta, GA, 2019.
[34] T. Solouki, M.R. Emmett, S. Guan, A.G. Marshall, Detection, number, and sequence location of sulfur-containing amino acids and disulfide bridges in peptides by ultrahigh-resolution MALDI FTICR mass spectrometry, Anal. Chem. 69 (1997) 1163–1168.
[35] F.W. McLafferty, Interpretation of Mass Spectra, third ed., University Science Books, Mill Valley, California, 1980.
[36] S.D.-H. Shi, C.L. Hendrickson, A.G. Marshall, Counting individual sulfur atoms in a protein by ultrahighresolution Fourier transform ion cyclotron resonance mass spectrometry: experimental resolution of isotopic fine structure in proteins, Proc. Natl. Acad. Sci. Unit. States Am. 95 (1998) 11532–11537.
[37] S. Kim, R.P. Rodgers, A.G. Marshall, Truly "exact" mass: elemental composition can be determined uniquely from molecular mass measurement at ~0.1mDa accuracy for molecules up to ~500Da, Int. J. Mass Spectrom. 251 (2006) 260–265.
[38] S. Kim, R.W. Kramer, P.G. Hatcher, Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van krevelen diagram, Anal. Chem. 75 (2003) 5336–5344.
[39] E.B. Kujawinski, M.D. Behn, Automated analysis of electrospray ionization fourier transform ion cyclotron resonance mass spectra of natural organic matter, Anal. Chem. 78 (2006) 4363–4373.
[40] J.C.L. Erve, M. Gu, Y. Wang, W. DeMaio, R.E. Talaat, Spectral accuracy of molecular ions in an LTQ/Orbitrap mass spectrometer and implications for elemental composition determination, J. Am. Soc. Mass Spectrom. 20 (2009) 2058–2069.
[41] A. Fievre, T. Solouki, A.G. Marshall, W.T. Cooper, High-resolution fourier transform ion cyclotron resonance mass spectrometry of humic and fulvic acids by laser desorption/ionization and electrospray ionization, Energy Fuel. 11 (1997) 554–560.
[42] J.A. Yergey, A general approach to calculating isotopic distributions for mass spectrometry, Int. J. Mass Spectrom. Ion Phys. 52 (1983) 337–349.
[43] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Inc., 1995.
[44] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 (1958) 386–408.
[45] G. Bebis, M. Georgiopoulos, Feed-forward neural networks, IEEE Potentials 13 (1994) 27–31.
[46] C.L. Giles, T. Maxwell, Learning, invariance, and generalization in high-order neural networks, Appl. Optic. 26 (1987) 4972–4978.
[47] R. Balestriero, Neural Decision Trees, 2017 arXiv:1702.07360 [cs, stat].
[48] F. Mühlberger, M. Saraji-Bozorgzad, M. Gonin, K. Fuhrer, R. Zimmermann, Compact ultrafast orthogonal acceleration time-of-flight mass spectrometer for on-line gas analysis by electron impact ionization and soft single photon ionization using an electron beam pumped rare gas excimer lamp as VUV-light source, Anal. Chem. 79 (2007) 8118–8124.
[49] M. Sud, E. Fahy, D. Cotter, A. Brown, E.A. Dennis, C.K. Glass, A.H. Merrill, R.C. Murphy, C.R.H. Raetz, D.W. Russell, S. Subramaniam, LMSD: LIPID MAPS structure database, Nucleic Acids Res. 35 (2007) D527–D532.
[50] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, A. Scalbert, HMDB 3.0—the human metabolome database in 2013, Nucleic Acids Res. 41 (2013) D801–D807.
[51] H. Liu, J. Zhang, H. Sun, C. Xu, Y. Zhu, H. Xie, The prediction of peptide charge states for electrospray ionization in mass spectrometry, Procedia Environmental Sciences 8 (2011) 483–491.
[52] M. Loos, C. Gerber, F. Corona, J. Hollender, H. Singer, Accelerated isotope fine structure calculation using pruned transition trees, Anal. Chem. 87 (2015) 5738–5744.
[53] M. Brantley, T. Solouki, Rapid and Non-targeted Detection of Chemical Substructures Using Feedforward Neural Networks, (Under Internal Review).
[54] L. Prechelt, Automatic early stopping using cross validation: quantifying the criteria, Neural Network. 11 (1998) 761–767.
[55] C.A. Hughey, C.L. Hendrickson, R.P. Rodgers, A.G. Marshall, K. Qian, Kendrick mass defect Spectrum: A compact visual analysis for ultrahigh-resolution broadband mass spectra, Anal. Chem. 73 (2001) 4676–4681.
[56] F.M. Fernández, TBI Lipidomics Dataset, Chorus Project, 2017.
[57] M.E. Pettit, F. Donnarumma, K.K. Murray, T. Solouki, Infrared laser ablation sampling coupled with data independent high resolution UPLC-IM-MS/MS for tissue analysis, Anal. Chim. Acta 1034 (2018) 102–109.
[58] J.R. Wiśniewski, A. Zougman, N. Nagaraj, M. Mann, Universal sample preparation method for proteome analysis, Nat. Methods 6 (2009) 359.
[59] V. Matyash, G. Liebisch, T.V. Kurzchalia, A. Shevchenko, D. Schwudke, Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics, J. Lipid Res. 49 (2008) 1137–1146.
[60] G. Paglia, P. Angel, J.P. Williams, K. Richardson, H.J. Olivos, J.W. Thompson, L. Menikarachchi, S. Lai, C. Walsh, A. Moseley, R.S. Plumb, D.F. Grant, B.O. Palsson, J. Langridge, S. Geromanos, G. Astarita, Ion mobility-derived collision cross section as an additional measure for lipid fingerprinting and identification, Anal. Chem. 87 (2015) 1137–1144.
[61] M.W. Senko, S.C. Beu, F.W. McLaffertycor, Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions, J. Am. Soc. Mass Spectrom. 6 (1995) 229–233.
[62] E.N. Nikolaev, R. Jertz, A. Grigoryev, G. Baykut, Fine structure in isotopic peak distributions measured using a dynamically harmonized fourier transform ion cyclotron resonance cell at 7 T, Anal. Chem. 84 (2012) 2275–2283.
[63] I.A. Popov, K. Nagornov, G.N. Vladimirov, Y.I. Kostyukevich, E.N. Nikolaev, Twelve million resolving power on 4.7 T fourier transform ion cyclotron resonance instrument with dynamically harmonized cell—observation of fine structure in peptide mass spectra, J. Am. Soc. Mass Spectrom. 25 (2014) 790–799.
[64] S.V. Stehman, Selecting and interpreting measures of thematic classification accuracy, Rem. Sens. Environ. 62 (1997) 77–89.
[65] E. Fahy, D. Cotter, M. Sud, S. Subramaniam, Lipid classification, structures and tools, Biochim. Biophys. Acta 1811 (2011) 637–647.
[66] J. Szulejko, S. Hall, M. Jackson, T. Solouki, Differentiation between pure cultures of Streptococcus pyogenes and Pseudomonas aeruginosa by FT-ICR-MS volatile analysis, Open Spectrosc. J. 3 (2009).
[67] M. Guilhaus, D. Selby, V. Mlynski, Orthogonal acceleration time-of-flight mass spectrometry, Mass Spectrom. Rev. 19 (2000) 65–107.
[68] A.N. Verentchikov, W. Ens, K.G. Standing, Reflecting time-of-flight mass spectrometer with an electrospray ion source and orthogonal extraction, Anal. Chem. 66 (1994) 126–133.