analytical chemistry

pubs.acs.org/ac Article

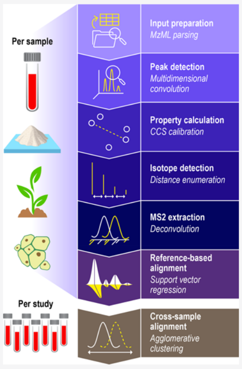
DEIMoS: An Open-Source Tool for Processing High-Dimensional Mass Spectrometry Data

Sean M. Colby, Christine H. Chang, Jessica L. Bade, Jamie R. Nunez, Madison R. Blumer, Daniel J. Orton, Kent J. Bloodsworth, Ernesto S. Nakayasu, Richard D. Smith, Yehia M. Ibrahim, Ryan S. Renslow,* and Thomas O. Metz*

Cite This: *Anal. Chem.* 2022, 94, 6130−6138 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: We present DEIMoS: Data Extraction for Integrated Multidimensional Spectrometry, a Python application programming interface (API) and command-line tool for high-dimensional mass spectrometry data analysis workflows that offers ease of development and access to efficient algorithmic implementations. Functionality includes feature detection, feature alignment, collision cross section (CCS) calibration, isotope detection, and MS/MS spectral deconvolution, with the output comprising detected features aligned across study samples and characterized by mass, CCS, tandem mass spectra, and isotopic signature. Notably, DEIMoS operates on *N*-dimensional data, largely agnostic to acquisition instrumentation; algorithm implementations simultaneously utilize all dimensions to (i) offer greater separation between features, thus improving detection sensitivity, (ii) increase alignment/feature matching confidence among data sets, and (iii) mitigate convolution artifacts in tandem mass spectra. We demonstrate DEIMoS with LC-IMS-MS/MS metabolomics data to illustrate the advantages of a multidimensional approach in each data processing step.

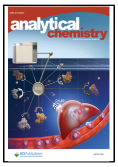


The ability to process raw instrument data reliably and accurately is critical to any molecular profiling assay. Though useful, commercial software solutions provided by vendors of mass spectrometry (MS) instrumentation lack flexibility required to rapidly adapt to evolving community needs. Demand for open-source, community-driven development has motivated researchers to pursue alternatives across instrument platforms, for example liquid or gas chromatography (LC or GC) and ion mobility spectrometry (IMS) coupled to mass spectrometry (MS), including tandem mass spectrometry (MS/MS). Software implementations also differ in their offered functionality: data input/output, multidimensional feature detection, alignment across samples, isotope detection, and deconvolution of MS/MS spectra. However, few available open-source, platform-agnostic solutions provide such core functionality for data of high-dimensionality, hindering the development and application of new instrumentation and analysis paradigms.

These limitations are predominantly tied to the specificity—and thus, relative inflexibility—of existing software and algorithm implementations. For example, LC or GC coupled to MS or MS/MS results in two primary feature dimensions, the retention time/index and MS mass-to-charge ratio (m/z), which is reflected in community software algorithms.[1−13] Existing feature detection algorithms are tailored to the underlying data: features are detected in one or two dimensions, and dimensions are often inflexibly constrained based on respective assumptions. In the short term, instrumentation advances[14−19] force platform-specific software to either ignore additional dimensions—for instance, summing across the least-distinguishing dimensions—or iteratively apply one- or two-dimensional algorithms.[20] Over time, wider instrument adoption engenders development to extend or modify existing algorithms to take full advantage of additional separation dimensions.[20,21] The problem is thus cyclical in nature: for software to adapt to technology, the technology must be mature and widely used, but for technology to mature and achieve widespread adoption, instrument data must be robustly analyzed by user-friendly software.

To overcome this paradox, instrument vendors have historically developed and supplied the software required to process the data, for example, Agilent MassHunter, Bruker MetaboScape, and Waters Progenesis QI. However, vendor offerings have their own limitations.[22] Because the underlying software is proprietary, details of underlying algorithm implementations are neither available to the public via open-source codebases nor sufficiently documented in publications. Moreover, vendor software is often tailored to a specific instrument and involves proprietary data formats, limiting one-to-one comparison of data across different instrument or

Received: November 18, 2021
Accepted: April 5, 2022
Published: April 17, 2022



https://doi.org/10.1021/acs.analchem.1c05017

vendor types. In some cases, existing software can also lack customizability; for instance, algorithm selection tends to be fixed to specific peak detection, alignment, and deconvolution implementations, unless additional options are explicitly implemented by the vendor. Thus, users are subject to the functionality provided by the vendor software, or must assemble multiple software solutions into one workflow.[3,23] Finally, many vendor solutions are automated only to a small degree, thus impeding reproducibility, and are not amenable to high-performance (HPC) or cloud computing.

As a result, the metabolomics community has worked to develop open-source solutions.[8,10,20,24−27] Each cover either some or all of the steps in a typical metabolomics workflow and are positioned to analyze GC-MS or LC-MS data (MS-DIAL[20] additionally handles some aspects of LC-IMS-MS data) and tandem MS. These software tools offer insight into best practices and algorithm implementations and serve as foundational references for our work. However, challenges remain in supporting data of arbitrary dimensionality, generalizing algorithmic implementations to operate in native dimensionality, offering flexibility and control over the analysis workflow, and scaling efficiently to computational resources.

To this end, we present the design and implementation of DEIMoS, or Data Extraction for Multidimensional Spectrometry, and include an initial evaluation on LC-IMS-MS/MS metabolomics data from analysis of blood plasma samples. DEIMoS's functionality is generalized through use of *N*-dimensional signal processing algorithms from the open-source, efficient, and widely used Python-based scientific computing packages NumPy[28] and SciPy.[29] Additionally, DEIMoS's design makes minimal assumptions about each underlying dimension. As a result, researchers may analyze GC-MS, LC-MS, IMS-MS, or LC-IMS-MS data, or another hypothetical MS-based platform, with or without MS/MS, using the same software with minimal reconfiguration. The underlying source code has also been written to account for hypothetical additional separation or analytical dimensions that may be introduced as instrumentation continues to advance (e.g., solid phase extraction[30] and associated chemical class-based separation, cryogenic infrared spectroscopy,[31] or multiplexed higher resolution ion mobility separations such as provided by structures for lossless ion manipulations, SLIM[32]). That is, calls to the application programming interface (API) and logic of the analysis may change, but the underlying source code can remain intact. This paradigm facilitates rapid advancement in metabolomics and introduces the potential to unify community efforts in informatics software development.

Furthermore, DEIMoS benefits from Python's rich existing ecosystem for scientific programming and offers even greater flexibility beyond the core API. DEIMoS's functionality is organized into several modules, each addressing one or more key data processing steps, including file input and output, peak detection, alignment, isotope detection, MS2 spectra extraction and deconvolution, and data subsetting operations. We describe each module relative to LC-IMS-MS/MS data, which represents higher dimensionality—and, by extension, complexity—among most current metabolomics analysis techniques. Data acquired on other platforms—for example, LC-MS(/MS), GC-MS(/MS)—require similar processing but in a lower dimensional space. Future algorithms and additional dimensions of data may be slotted in easily.

We architected DEIMoS to adhere to software development best practices,[33] including installation through Anaconda[34] or PyPI,[35] in-line documentation via docstrings and aggregation via *Sphinx*,[36] unit test implementations with *pytest*[37] coupled with continuous integration and static code coverage analysis, and version control with *Git*.[38] DEIMoS is open-source and freely available online at github.com/pnnl/deimos, and community contributions via pull request are welcome. Documentation, including a user guide, API reference, examples, and contribution instructions are available at deimos.readthedocs.io.

■ METHODS

Experimental Methods. To demonstrate DEIMoS, we examined LC-IMS-MS/MS data from a large study of human plasma samples consisting of 40 quality control (QC) samples from the NIST Standard Reference Material 1950[39] and 112 study samples. An internal standard mixture was added prior to extraction, its composition listed in the SI. Each sample was spiked with 50 μL of a solution of the internal standards at 0.166 mg/mL in water. Metabolites and lipids were extracted with concomitant protein precipitation using the Matyash protocol[40] described previously.[41] The metabolite layer was removed and dried in vacuo. Lipid and protein layers were not analyzed.

An Agilent 1260 Infinity II high flow liquid chromatography system (San Jose, CA) equipped with a Vial Sampler and Binary Pump was used to inject and chromatographically separate samples prior to introduction to the ion mobility spectrometry-mass spectrometry instrument. A steady flow rate of 0.300 mL/min was delivered through a Millipore-Sigma (Burlington, MA) SeQuant Zic-pHILIC column (15 cm length, 2.1 mm inner diameter, packed with 5 μm particles). A corresponding guard column of the same packing material was also used. Mobile phases consisted of (A) 20 μM ammonium acetate in water and (B) 100% acetonitrile with the following gradient profile (min, %B): 0, 90; 4, 90; 12, 20; 13, 10; 15, 10; 17, 90. An Agilent 1100 Column Heater was used with a static temperature of 45 °C.

Ion mobility spectrometry-tandem mass spectrometry analysis was performed using an Agilent 6560 Ion Mobility LC/Q-TOF system. Spectra were acquired separately in both positive and negative ionization modes. Data were collected in the mass range of 50−1700 *m*/*z*. Ionization was accomplished using a Dual AJS ESI source, with gas temperature set to 325 °C, drying gas set to 5 L/min, nebulizer set to 30 psi, sheath gas temperature set to 275 °C, sheath gas flow of 11 L/min, VCap set to 2500 V, nozzle voltage set to 2000 V, and the fragmentor set to 400 V. For the ion mobility separations, the trap fill time was 30 000 μs and released for 300 μs. Frame rate was 1 frame/s, 19 IM transitions/frame, and max drift time set to 50 ms. Fixed collision energies were employed at 10, 20, and 40 eV on alternating frames. Data were collected for 22 min immediately following the injection of the sample.

To communicate generalizability to other instrument platforms and/or data types, we acquired data from a Bruker timsTOF Pro and a Waters UPLC i-Class coupled to a Synapt G2Si. The former was obtained from the Mass Spectrometry Interactive Virtual Environment[42] (MassIVE, accession MSV000088020), the latter from Metabolights[43] (accession MTBLS812). The Bruker instrument implements trapped ion mobility, resulting in measurements of inverse reduced mobility as opposed to drift time; the Waters instrument implements traveling wave ion mobility.

Overview. DEIMoS was developed as an instrument-independent, high-dimensional metabolomics data analysis tool, and design choices reflect this philosophy. DEIMoS is written in Python, prioritizing user productivity, and ease of development and use. While high-level interpreted languages, such as Python, often suffer from reduced computational efficiency, many popular scientific Python libraries, such as NumPy[28] and SciPy,[29] wrap C or Fortran code and are, thus, highly optimized. In addition, Python is ubiquitous across the sciences and in industry, user-friendly, and largely agnostic to client platform (Windows, macOS, Linux). As a result, Python boasts a large, active, and continually growing community in the sciences, positioning Python-based software for wide adoption both by users and collaborative developers.[44] DEIMoS is one of few Python-based offerings for metabolomics data processing uniquely offering support for data of any dimension, algorithmic implementations that operate in native dimensionality, flexibility and control over the analysis workflow, and efficient scaling to computational resources. An overview of functionality is depicted in Figure 1. All computation was performed on AMD EPYC 7502 CPUs with 4 GB of memory per core.

File Input/Output. To accommodate disparate instrument types and manufacturers (e.g., Bruker, Waters, Thermo, Agilent), DEIMoS operates under the assumption that input data are in an open, standard format. As of this publication, the accepted file format for DEIMoS is mzML,[45] which contains metadata, separation, and spectrometry data that reproduce the contents of vendor formats. Conversion to mzML from several other formats can be performed using the free and open-source ProteoWizard *msconvert* utility.[24] By default, DEIMoS exports a lightweight, data frame-based representation in Hierarchical Data Format version 5 (HDF5) file format.[46] Additionally, adaptors are included to support exporting to delimited text (e.g., CSV), Mascot Generic Format (MGF), and mzML for downstream use with other tools (e.g., MAME,[47] LIQUID,[48] GNPS[49]).

Feature Detection. Feature detection, also referred to as peak detection, is the process by which local maxima that fulfill certain criteria (such as sufficient signal-to-noise ratio) are located in the signal acquired by a given analytical instrument. This process results in "features" associated with the analysis of molecular analytes from the sample under study or from chemical, instrument, or random noise. Typically, feature detection involves a mass dimension (m/z), as well as one or more separation dimensions, the latter offering distinction among isobaric/isotopic features.

DEIMoS implements an N-dimensional maximum filter from *scipy.ndimage* that convolves the instrument signal with a structuring element, also known as a kernel, and compares the result against the input array to identify local maxima as candidate features or peaks. We discuss what qualifies as a dimension in the SI. Additional filters, including integral and average intensity, kurtosis, skew, etc., can be applied to yield statistics for later downselection. To provide additional confidence in detected features, we required that a given feature be observed across analytical triplicates.

Key to this process is the selection of kernel size, which can vary by instrument, data set, and even compound. For example, in LC-IMS-MS/MS data, peak width increases with increasing m/z and drift time, and also varies in retention time. Ideally, the kernel would be the same size as the N-dimensional peak (i.e., wavelets[1,5,12,50]), though computational efficiency considerations for high-dimensional data currently limit the ability to dynamically adjust kernel size. Thus, the selected kernel size should be representative of likely features of interest. In some scenarios, dynamic kernel size may be appropriate, per the kernel selection discussion in the SI.

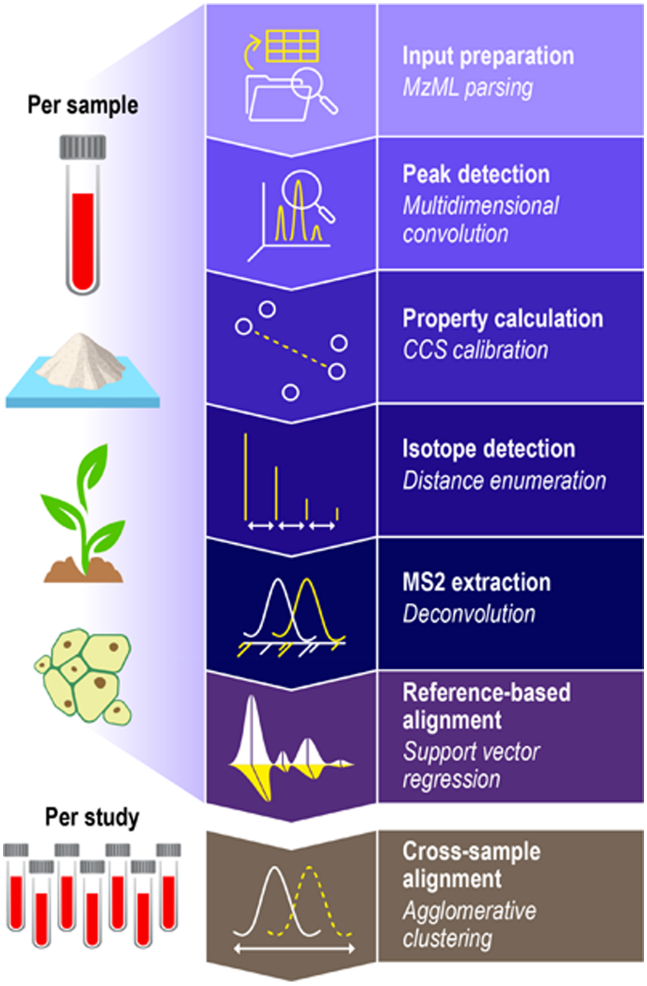


Figure 1. Functionality overview. High level overview of available DEIMoS functionality, with operations delineated as "per sample" versus "per study". That is, the former operations are performed for each instrument acquisition, whereas the latter is performed among all data from comparable samples acquired. Functionality and underlying methods are easily extended or modified.

While we recommend processing the data in its native dimensionality, DEIMoS's algorithms are flexible and can detect features in iterative subspaces, for example 2D followed by 1D, 1D followed by 2D, or successive 1D. We used the same parameters per dimension to evaluate feature detection in all dimensional permutations for LC-IMS-MS data. Features were only kept if they appeared across all three analytical replicates. To compare methods, we (i) compared feature coordinates directly and (ii) used tolerances of ±20 ppm, ± 1.5%, and ±0.3 min for m/z, drift time, and retention time, respectively, based on peak dimensions determined during kernel size selection (Bruker and Waters values differ here, as reported in the SI). We used relative tolerances, such as parts-per-million and percent for m/z and drift time, respectively, because unlike in the retention time dimension, peak widths in *m*/*z* and drift time varied with mass (Figure S1). This analysis was performed for all samples from the study, averaged per ionization mode.

Alignment. Alignment is the process by which feature coordinates across samples are adjusted to account for instrument variation (drift, calibration, etc.) such that matching features are aligned to adjust for small differences in coordinates. To perform alignment, we first constructed a model for each dimension of a sample by putatively matching detected features against an in-study reference sample, minimizing the residual, and subsequently applying the fit transform. Next, we matched corresponding features across data sets within a user defined tolerance. We refer to the former as "reference-based alignment" and the latter as "cross-sample alignment".

For reference-based alignment, we defined corresponding features between two samples based on minimum distance in the dimension of interest and selected tolerances to accommodate potentially complex nonlinear relationships. We suggest visualizing putative matches with multiple tolerance selections. Once features were matched, we modeled the relationship between samples using support vector regression (SVR) as implemented in *scikit-learn*.[51] While SVR was selected here for its broad applicability to both linear and nonlinear alignment, many approaches have been successfully developed in this space,[2,13,22,52−55] and SVR is not necessarily superior.

Many existing algorithms and implementations can perform cross-sample alignment.[8,13,55] We initially explored use of a modified version of the "join align" method from MZmine[8] but ultimately arrived at an agglomerative clustering-based approach. Though similar with respect to resulting alignment, the agglomerative clustering-based approach was more amenable to processing many samples simultaneously. Additional details on the agglomerative clustering implementation can be found in the SI.

By default, alignment considers all features detected among data sets, though users may design more complicated and restrictive workflows using the DEIMoS API. For example, users may choose to only align features that appear across some number of replicates or exclude features that appear in blank samples.

To demonstrate alignment functionality, we analyzed all acquired data files (N = 912, 54, and 21 for Agilent, Bruker, and Waters data, respectively). First, we performed alignment across analytical replicates and only kept features appearing in triplicate. Next, we performed alignment across samples in both positive and negative electrospray ionization (ESI) modes, as available. Samples were aligned by agglomerative clustering method with maximum linkage distance tolerances in each dimension of ±20 ppm, ± 1.5%, and ±0.3 min for *m*/*z*, drift time, and retention time, respectively. Tolerances for Bruker and Waters values differ, as reported in the SI.

MS2 Extraction. With MS1 features of interest determined by peak detection, corresponding tandem mass spectra, if available, must be extracted and assigned to the MS1 parent ion feature. For data independent acquisition, we use non-*m*/*z* dimensions to assign fragments; for instance, drift time and retention time are used to match fragments in LC-IMS-MS/MS. These additional separations enable better attribution of MS2 ions to parent ions, a form of deconvolution inherent in the acquisition, but convolution artifacts can still occur.

Explicit, algorithmic deconvolution[20,25,56−58] has been implemented in DEIMoS such that MS1 and MS2 features overlapping in non-*m*/*z* separation dimensions are disambiguated to minimize false assignments. In this form of deconvolution, similar to the approach in Yin et al.,[58] the profiles of non-*m*/*z* separation dimensions are used to identify only those ions in the MS2 with distributions that correspond to the precursor ion distribution. This technique simultaneously excludes MS2 ions arising from noise or chemical background, while also attributing MS2 ions only to precursor ions with similar separation distributions. Correspondence is determined by cosine similarity, producing a value between 0 and 100 for each separation dimension for all MS1: MS2 pairings. The user may then filter putative matches by this value, for example considering only those above some tolerance in one or more of the separation dimensions.

Collision Cross Section Calibration. To yield collision cross section (CCS) from IMS arrival time, a calibration must be performed using a standard tune mix containing compounds of known CCS. Drift times, or analogous measurement, such as inverse reduced mobility in TIMS, are reported by the instrument and calibrated against the known CCS values to yield calibration coefficients *beta* and *tfix*. For drift tube and trapped IMS, the single-field calibration equation detailed in Stow et al. and as implemented in Lee et al. was used.[59,60] For traveling wave IMS, the relationship between measurement and CCS was first linearized by the natural logarithm, then fit by linear regression. DEIMoS performs this calibration given arrival times (or analogous measurement), known CCS values, *m*/*z*, and nominal charge of each calibrant. Correlation coefficient and sum of residuals are reported to characterize goodness of fit. Users may also supply *beta* and *tfix* directly.

Extracted Ion Approach. DEIMoS can locate features based on extracted ion chromatograms (XIC), mobilograms (XIM), or multidimensional analogs. Here, a specified *m*/*z* of interest is supplied and the feature of maximal intensity in the remaining dimensions is returned. This technique is useful, for example, when detecting an internal standard that has been spiked into a sample or when single or mixtures of pure compounds are analyzed. Adduct *m*/*z* were calculated using the mass spectrometry adduct calculator (MSAC).[61] We recommend multidimensional representations for targeted feature detection. Further details are included in the SI.

Isotope Detection. Isotopologues, or molecules that differ only in their isotopic composition, are common in mass spectrometry analyses. In many analysis workflows, isotopologues are used to down select the total feature list to include only the most abundant feature, as well as to glean ion charge state and provide further evidence for identification by way of a detected isotopic signature. Details of the DEIMoS implementation, as well as examples of isotopic signatures for a singly charged feature, a multiply charged feature, and overlapping features, are available in the SI.

Automation. While DEIMoS functionality is implemented as a Python API, a typical workflow has been implemented using the Snakemake[62] workflow management system[62] and made accessible via command line interface (CLI). Users need only modify a configuration file and interact with a CLI to process input mzML files into output feature coordinates and extracted MS2. Moreover, Snakemake can automatically handle scaling to HPC and cloud resources, enabling the high-throughput processing of numerous samples. A graphical user interface (GUI) is currently in development to facilitate accessibility of DEIMoS to those without programming experience and will be reported in a subsequent manuscript.

■ RESULTS AND DISCUSSION

In total, sample acquisition resulted in 912 data files for 112 study samples, 40 quality control samples, and 10 blanks, each in positive and negative ionization modes and collected at 3 collision energies (10, 20, and 40 eV): cumulatively 1.1 terabytes. These data were processed by DEIMoS using feature detection, alignment, CCS calibration, and MS2 extraction by deconvolution.

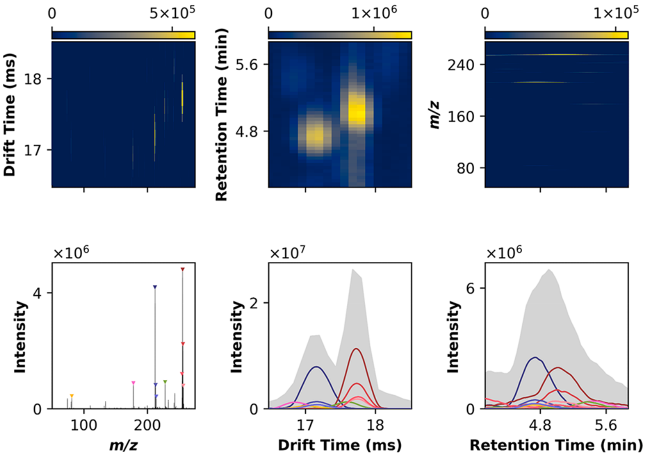


Figure 2. Multidimensional peak detection. Peak detection involves convolving the input signal in N dimensions (here, in LC-IMS-MS, 3D) with a maximum filter. The input and maximum filtered arrays are then compared point-by-point and, where equal, a local maximum is indicated. While the data is collected in 3D, this approach is best visualized in 2D and 1D projections, capturing all lower dimensional representations of the underlying 3D data. Note a well-defined peak in a given 2D view may or may not correspond to a true 3D apex, or can be the product of multiple underlying features. It is, thus, important to interpret the 1D projections carefully. For this subset of the data, the top 10 most intense local maxima are shown, colored by m/z, with similar m/z (i.e., isopologues) sharing hues.

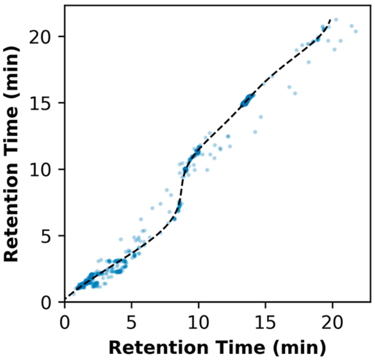


Figure 3. Nonlinear alignment by support vector regression. Support vector regression (SVR) was evaluated here on the retention time dimension between 2 illustrative samples described by a nonlinear, "S-shaped" relationship in retention time. To model this relationship, a radial basis function (RBF) kernel was selected. Measurements between samples varied negligibly in drift time and m/z, and thus alignment was only necessary in retention time. An example involving samples with a linear relationship in retention time is included in Figure S7.

Acquired Bruker data was comprised of 54 data files for 9 study samples in triplicate, spanning 2 ionization modes: cumulatively 83 gigabytes. The Waters data consisted of 21 data files for 6 study samples in triplicate and 3 blanks, each in positive ionization mode: cumulatively 16 gigabytes. These data were processed similarly, though no CCS calibration nor MS2 deconvolution was performed for the Bruker data, in that neither CCS calibration data nor MS2 was available.

Feature Detection. Feature detection for a subset of a single sample has been visualized in Figure 2. The selected region illustrates the inherent convolution of MS1 features leads to overlap in both drift and retention time, resulting in several putative precursor ions for MS2 assignment. The requirement of explicit deconvolution to appropriately attribute ions in the MS2 spectra becomes apparent, as features are not sufficiently resolved by drift and retention time coordinates. Representative features extracted from the Bruker and Waters data are also included in Figures S2 and S3.

For all acquired samples, we compared permutations of the possible feature detection modes (3D, 2D followed by 1D, 1D followed by 2D, and iterative 1D). Notably, LC-IMS-MS/MS data exist in a 3D space; thus, the underlying features are also represented in 3D. Iterative feature detection in lower-dimensional projections simplifies the resulting data structure by summing along nonprojection axes, potentially introducing artifacts. In practice, some projections, such as m/z versus drift time, are affected less significantly than others, such as drift time versus retention time, the latter suffering significant information loss when summing along the m/z axis.

We anticipated that processing the data in all 3 dimensions simultaneously would result in the greatest separation among features to better isolate the local maxima. That is, given the same feature detection tolerances across methods, 3D feature detection would theoretically afford the least overlap and, by extension, greatest number of features. Per Figure S4, a contrary result was, thus, surprising. However, the coordinates of the features detected by lower dimensional projections are not always congruent with the 3D approach (Figure S5). This signals that the projections along the various data axes, whether 1D or 2D, aggregate signal to the point of, in some cases, losing the underlying feature defined in 3D − the sum operation along a given axis "merges" previously separated features, skewing the coordinate in that dimension.

However, this phenomenon is pronounced to varying degree among methods: the most comparable technique, as implemented by MS-DIAL—m/z versus RT followed by DT—results in poor agreement when considering strict tolerances (only ∼13% intersection in both positive and negative mode), but intersection increases substantially (to ∼90%) when imposing the same tolerances that would be used in cross-sample alignment. That is, tolerances that would result in the combining of those features anyway. In this case, the lower dimensional projections result in slight differences in feature coordinates, but in practical application would be treated as "same". The difference resulted in slightly less accurate characterization of feature coordinates, for example exact m/z, drift time (and by extension CCS), and retention time, where different, deviated by an average of 5.8 ppm, 0.2%, and 0.05 min for m/z, drift time, and retention time, respectively, when comparing 3D processing to m/z versus RT followed by DT. Additional comparisons are depicted in

https://doi.org/10.1021/acs.analchem.1c05017

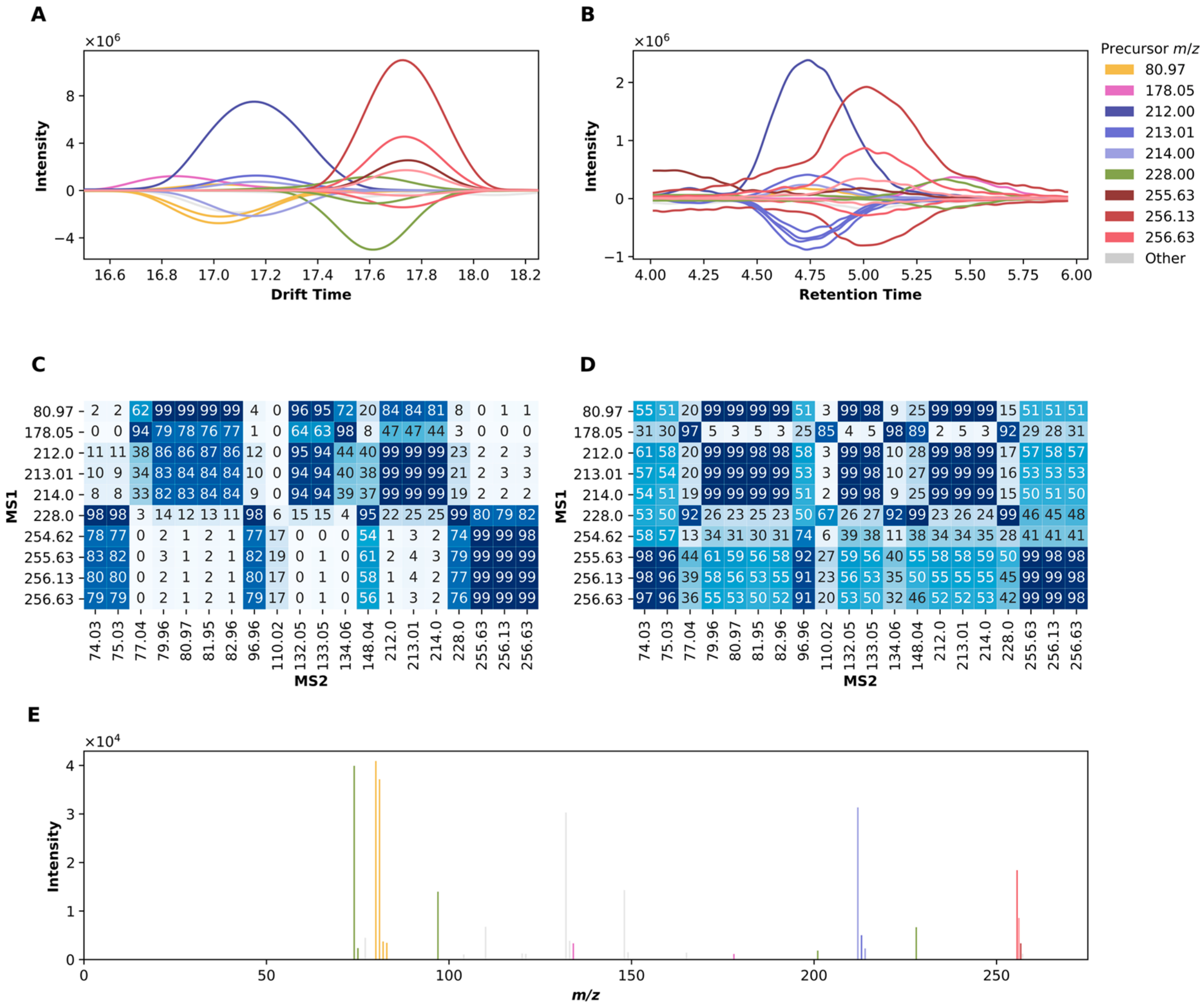


Figure 4. MS2 deconvolution. The MS2 spectra belonging to the MS1 features highlighted in Figure 2 were algorithmically deconvolved. The profiles of the MS1 features are indicated by respective colors, plotted along the positive *y*-axis for drift time (A) and retention time (B). These are accompanied by corresponding MS2 profiles, plotted along the negative *y*-axis, colored according to the closest matching MS1 profile by cosine similarity. Panels C and D show the pairwise cosine similarity of MS1 and MS2 profiles for drift and retention time, respectively. In panel E, ions in the MS2 spectra are colored according to the closest matching MS1 drift time profile. As in Figure 2, only the 10 most intense MS1 features were explicitly colored; the ions in the MS2 spectra corresponding to remaining MS1 features, or not sufficiently similar to an MS1 precursor, are indicated in gray. Note that disambiguating among isotopologues is not possible here; thus, the difference in color among isotopic groupings is largely superficial.

Figure S6. Critically, the order of peak picking operations had a large impact on the number and composition of the features detected (e.g., comparing 1D−2D to 2D−1D operations, as well as the consecutive 1D operations).

A key advantage of feature detection in native dimensionality is that computation time does not scale with the number of features (Figure S4). Peak detection in native dimensionality required 0.54 ± 0.11 core-hours and 0.57 ± 0.11 core-hours per data file for positive and negative ionization modes, respectively (mean ± standard deviation). Cumulatively, 373.74 core-hours for all 912 data files.

For the Bruker data, because the data were collected at higher resolution (1 728 842, 937, and 4769 unique values for *m*/*z*, inverse reduced mobility, and retention time, respectively; collectively 165 times the Agilent resolution), processing time for peak detection was much longer at 12.20 ± 1.35 core-hours and 35.83 ± 15.23 core-hours per data file for positive and negative ionization modes, respectively (mean ± standard deviation). The Waters data, which was collected at lower resolution (79 863, 200, and 707 unique values for *m*/*z*, drift time, and retention time, respectively; collectively 0.25 times the Agilent resolution), required less processing time for peak detection: 0.33 ± 0.03 core-hours for positive ionization mode (negative mode not collected).

The longer processing times for the negative ionization mode data were due to signal being observed over a larger range of unique *m*/*z*, drift (or analogous measurement), and retention times. The array partitioning leveraged for multicore computation split along the *m*/*z* dimension. For each partition, computation scaled with the number of observed measurements spanning remaining dimensions. Computation also scaled with number of partitions, in that empty partitions were not processed. To ameliorate, a nominal intensity threshold may be applied to decrease signal span of the 3D space. Said threshold must be selected at or below the noise floor in the data.

Alignment. We examined results from various SVR kernels and found that the linear kernel achieved satisfactory results when the instrument misalignment could be corrected by linear regression, whereas the radial basis function (RBF) kernel was able to account for nonlinear relationships. For instance, the RBF SVR model could match features when samples were run consecutively on a degrading LC column.

We selected example data sets to highlight both S-shaped (Figure 3) and linear (Figure S7) relationships in prealignment retention time and illustrate the flexibility of the SVR-based approach. In these data, m/z and drift time were already sufficiently aligned; as such, plots for m/z and drift time were omitted. The potential difference realized by kernel selection motivates visual confirmation of the alignment relationship between samples: an RBF kernel applied to linearly related samples would be considered overfit, whereas a linear kernel applied to a nonlinear case would achieve poor alignment.

Subsequent cross-sample alignment by agglomerative clustering resulted in 11 698 and 14 784 features for positive and negative mode appearing across each of 3 analytical replicates, excluding features that appeared in all blank samples. Though tolerances were selected according to expected variance across samples, agglomerative clustering resulted in tighter linkages (Figure S8), indicating intersample variance is lower than typical peak width. This variance reaches near-maximal values at 15 ppm in m/z, 1% in drift time, and 0.2 min in retention. Thus, exploratory analysis of the data enables users to characterize interfeature and intersample variance to set appropriate tolerances relative to the source of maximum variance, though agglomerative clustering is relatively forgiving so long as tolerances are not too small.

With respect to computation time, reference-based alignment required on the order of milliseconds per file. For cross-sample alignment, each set of 3 analytical replicates ($N = 304$) required 39.89 ± 6.59 and 44.90 ± 5.53 core-seconds for positive and negative mode, respectively (mean ± standard deviation); cumulatively 3.58 core-hours. Per-file processing times did not significantly differ for the Bruker and Waters data.

Aligning those features that appeared in each analytical replicate required 0.66 and 0.94 core-hours for positive and negative mode, respectively; cumulatively 1.60 core-hours. Because of the relatively smaller number of samples, this step only required seconds for the Bruker and Waters data.

MS2 extraction. For the MS1 features shown in Figure 2, we performed deconvolution to putatively assign ions in the MS2 spectra to corresponding ions in the MS1 spectra. The utility of deconvolution is highlighted in Figure 4. A window-based approach yields MS2 spectra that, in the worst case, erroneously include all ions for all MS1 precursors to give identical MS2 spectra or, in the best case, results in only two distinct spectra among precursors. The limitation here is visualized by the plot of drift time versus retention time, where only two overlapping peaks emerge. Naïve assignment would thus yield convolved spectra, the degree of convolution depending on window selection.

Notably, deconvolution results may differ substantially if employing the cosine similarity scores of drift versus retention time, as false positives can occur if using retention time alone. For example, the masses between 77 and 83 Da would be assigned to the precursor with m/z 212 using retention time, whereas they are sufficiently separated in drift time (Figure 4).

Computation time required to deconvolve MS2 spectra was a function of number of distinct non-m/z separation dimension populations, as determined by agglomerative clustering, and the size of each such population. Figure 4 shows one such population. In this analysis, each data file required 1.18 ± 0.28 core-hours and 2.27 ± 0.31 (mean ± standard deviation) for positive and negative mode, respectively; cumulatively 1579.07 core-hours. Per-file processing times did not significantly differ for the Waters data (MS2 not collected for the Bruker data).

■ CONCLUSION

Metabolomics and exposomics data processing tools offer immense value for diagnosis of disease, evaluation of environmental exposures, and discovery of novel molecules. However, few open-source solutions are currently positioned to fully leverage the latest instrumentation. Importantly, though demonstrated for LC-IMS-MS/MS data, DEIMoS's architecture supports extension to other measurement modalities, such as cryogenic infrared spectroscopy, minimizing development barriers as instrumentation evolves. Further, all development has been accomplished using design principles necessary for the long-term success for metabolomics data: format interoperability, workflow flexibility, open-source software implementation, community development, and reproducibility.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.1c05017.

Commentary on select terminology and additional methods, results, and discussion, including supporting figures (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Ryan S. Renslow – *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0002-3969-5570; Email: ryan.renslow@pnnl.gov

Thomas O. Metz – *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0001-6049-3968; Email: thomas.metz@pnnl.gov

Authors

Sean M. Colby – *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Christine H. Chang – *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Jessica L. Bade – *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Jamie R. Nunez – *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Madison R. Blumer – *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Daniel J. Orton – *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Kent J. Bloodsworth − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States*

Ernesto S. Nakayasu − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0002-4056-2695

Richard D. Smith − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0002-2381-2349

Yehia M. Ibrahim − *Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington 99352, United States;* orcid.org/0000-0001-6085-193X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.1c05017

Notes

The authors declare no competing financial interest.
The raw LC-IMS-MS/MS data files from this study in mzML format were deposited and are publicly available at the MassIVE repository (MSV000088849).

■ ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health, National Institute of Environmental Health Sciences Grant U2CES030170. Additional support was provided by the Pacific Northwest National Laboratory (PNNL), Laboratory Directed Research and Development Program, and is a contribution of the Biomedical Resilience & Readiness in Adverse Operating Environments (BRAVE) Agile project. The authors wish to thank graphics artist Mr. Nathan Johnson at PNNL for support in creating Figure 1. LC-IMS-MS/MS measurements were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the U.S. Department of Energy Office of Biological and Environmental Research and located on the campus of Pacific Northwest National Laboratory (PNNL) in Richland, Washington. PNNL is a multiprogram national laboratory operated by Battelle for the DOE under Contract DE-AC05-76RLO 1830.

■ REFERENCES

(1) Wang, P.; Yang, P.; Arthur, J.; Yang, J. Y. H. *Bioinformatics* **2010**, *26* (18), 2242−2249.

(2) Katajamaa, M.; Orešič, M. *J. Chromatogr. A* **2007**, *1158* (1), 318−328.

(3) Kiefer, P.; Schmitt, U.; Vorholt, J. A. *Bioinformatics* **2013**, *29* (7), 963−964.

(4) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinformatics* **2008**, *9* (1), 504.

(5) Du, P.; Kibbe, W. A.; Lin, S. M. *Bioinformatics* **2006**, *22* (17), 2059−2065.

(6) DeFelice, B. C.; Mehta, S. S.; Samra, S.; Čajka, T.; Wancewicz, B.; Fahrmann, J. F.; Fiehn, O. *Anal. Chem.* **2017**, *89* (6), 3250−3255.

(7) Broeckling, C. D.; Reddy, I. R.; Duran, A. L.; Zhao, X.; Sumner, L. W. *Anal. Chem.* **2006**, *78* (13), 4334−4341.

(8) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. *BMC Bioinformatics* **2010**, *11* (1), 395.

(9) Hastings, C. A.; Norton, S. M.; Roy, S. *Rapid Commun. Mass Spectrom.* **2002**, *16* (5), 462−467.

(10) Röst, H. L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; Liang, X.; Nahnsen, S.; Nilse, L.; Pfeuffer, J.; Rosenberger, G.; Rurik, M.; Schmitt, U.; Veit, J.; Walzer, M.; Wojnar, D.; Wolski, W. E.; Schilling, O.; Choudhary, J. S.; Malmström, L.; Aebersold, R.; Reinert, K.; Kohlbacher, O. *Nat. Methods* **2016**, *13* (9), 741−748.

(11) Monroe, M. E.; Tolić, N.; Jaitly, N.; Shaw, J. L.; Adkins, J. N.; Smith, R. D. *Bioinformatics* **2007**, *23* (15), 2021−2023.

(12) French, W. R.; Zimmerman, L. J.; Schilling, B.; Gibson, B. W.; Miller, C. A.; Townsend, R. R.; Sherrod, S. D.; Goodwin, C. R.; McLean, J. A.; Tabb, D. L. *J. Proteome Res.* **2015**, *14* (2), 1299−1307.

(13) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (3), 779−787.

(14) Blaženović, I.; Shen, T.; Mehta, S. S.; Kind, T.; Ji, J.; Piparo, M.; Cacciola, F.; Mondello, L.; Fiehn, O. *Anal. Chem.* **2018**, *90* (18), 10758−10764.

(15) Dodds, J. N.; Baker, E. S. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (11), 2185−2195.

(16) Lanucara, F.; Holman, S. W.; Gray, C. J.; Eyers, C. E. *Nat. Chem.* **2014**, *6* (4), 281−294.

(17) May, J. C.; McLean, J. A. *Anal. Chem.* **2015**, *87* (3), 1422−1436.

(18) Metz, T. O.; Baker, E. S.; Schymanski, E. L.; Renslow, R. S.; Thomas, D. G.; Causon, T. J.; Webb, I. K.; Hann, S.; Smith, R. D.; Teeguarden, J. G. *Bioanalysis* **2017**, *9* (1), 81−98.

(19) Paglia, G.; Williams, J. P.; Menikarachchi, L.; Thompson, J. W.; Tyldesley-Worster, R.; Halldórsson, S.; Rolfsson, O.; Moseley, A.; Grant, D.; Langridge, J.; Palsson, B. O.; Astarita, G. *Anal. Chem.* **2014**, *86* (8), 3985−3993.

(20) Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; Higashi, Y.; Okazaki, Y.; Zhou, Z.; Zhu, Z.-J.; Koelmel, J.; Cajka, T.; Fiehn, O.; Saito, K.; Arita, M.; Arita, M. *Nat. Biotechnol.* **2020**, *38* (10), 1159−1163.

(21) Crowell, K. L.; Slysz, G. W.; Baker, E. S.; LaMarche, B. L.; Monroe, M. E.; Ibrahim, Y. M.; Payne, S. H.; Anderson, G. A.; Smith, R. D. *Bioinformatics* **2013**, *29* (21), 2804−2805.

(22) Fernández-Ochoa, Á.; Quirantes-Piné, R.; Borrás-Linares, I.; de la Cádiz-Gurrea, M.; PRECISESADS Clinical Consortium; Alarcón Riquelme, M. E.; Brunius, C.; Segura-Carretero, A. *Metabolites* **2020**, *10* (1), 28.

(23) Aiche, S.; Sachsenberg, T.; Kenar, E.; Walzer, M.; Wiswedel, B.; Kristl, T.; Boyles, M.; Duschl, A.; Huber, C. G.; Berthold, M. R.; Reinert, K.; Kohlbacher, O. *PROTEOMICS* **2015**, *15* (8), 1443−1447.

(24) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics* **2008**, *24* (21), 2534−2536.

(25) Benton, H. P.; Wong, D. M.; Trauger, S. A.; Siuzdak, G. *Anal. Chem.* **2008**, *80* (16), 6382−6389.

(26) Melamud, E.; Vastag, L.; Rabinowitz, J. D. *Anal. Chem.* **2010**, *82* (23), 9818−9826.

(27) Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J. *Nucleic Acids Res.* **2018**, *46* (W1), W486−W494.

(28) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. *Nature* **2020**, *585* (7825), 357−362.

(29) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; et al. *Nat. Methods* **2020**, *17* (3), 261−272.

(30) Zhang, X.; Romm, M.; Zheng, X.; Zink, E. M.; Kim, Y.-M.; Burnum-Johnson, K. E.; Orton, D. J.; Apffel, A.; Ibrahim, Y. M.; Monroe, M. E.; Moore, R. J.; Smith, J. N.; Ma, J.; Renslow, R. S.; Thomas, D. G.; Blackwell, A. E.; Swinford, G.; Sausen, J.; Kurulugama, R. T.; Eno, N.; Darland, E.; Stafford, G.; Fjeldsted, J.;

Metz, T. O.; Teeguarden, J. G.; Smith, R. D.; Baker, E. S. *Clin. Mass Spectrom. Mar Calif* **2016**, *2*, 1−10.

(31) Warnke, S.; Ben Faleh, A.; Pellegrinelli, R. P.; Yalovenko, N.; Rizzo, T. R. *Faraday Discuss.* **2019**, *217* (0), 114−125.

(32) Deng, L.; Ibrahim, Y. M.; Hamid, A. M.; Garimella, S. V. B.; Webb, I. K.; Zheng, X.; Prost, S. A.; Sandoval, J. A.; Norheim, R. V.; Anderson, G. A.; Tolmachev, A. V.; Baker, E. S.; Smith, R. D. *Anal. Chem.* **2016**, *88* (18), 8957−8964.

(33) Chang, H.-Y.; Colby, S. M.; Du, X.; Gomez, J. D.; Helf, M. J.; Kechris, K.; Kirkpatrick, C. R.; Li, S.; Patti, G. J.; Renslow, R. S.; Subramaniam, S.; Verma, M.; Xia, J.; Young, J. D. *Anal. Chem.* **2021**, *93* (4), 1912−1923.

(34) *Anaconda*. https://www.anaconda.com/ (accessed 2020-12-10).

(35) *PyPI*. https://pypi.org/ (accessed 2020-12-10).

(36) *Sphinx*. https://www.sphinx-doc.org/en/master/ (accessed 2020-12-10).

(37) *pytest*. https://docs.pytest.org/en/stable/ (accessed 2020-12-10).

(38) *Git*. https://git-scm.com/ (accessed 2020-12-10).

(39) Simón-Manso, Y.; Lowenthal, M. S.; Kilpatrick, L. E.; Sampson, M. L.; Telu, K. H.; Rudnick, P. A.; Mallard, W. G.; Bearden, D. W.; Schock, T. B.; Tchekhovskoi, D. V.; Blonder, N.; Yan, X.; Liang, Y.; Zheng, Y.; Wallace, W. E.; Neta, P.; Phinney, K. W.; Remaley, A. T.; Stein, S. E. *Anal. Chem.* **2013**, *85* (24), 11725−11731.

(40) Matyash, V.; Liebisch, G.; Kurzchalia, T. V.; Shevchenko, A.; Schwudke, D. *J. Lipid Res.* **2008**, *49* (5), 1137−1146.

(41) Lee, D. Y.; Kind, T.; Yoon, Y.-R.; Fiehn, O.; Liu, K.-H. *Anal. Bioanal. Chem.* **2014**, *406* (28), 7275−7286.

(42) *MassIVE*. https://massive.ucsd.edu/ (accessed 2022-03-01).

(43) Haug, K.; Cochrane, K.; Nainala, V. C.; Williams, M.; Chang, J.; Jayaseelan, K. V.; O'Donovan, C. *Nucleic Acids Res.* **2019**, *48* (D1), D440−D444.

(44) Pittard, W. S.; Li, S. The Essential Toolbox of Data Science: Python, R, Git, and Docker. In *Computational Methods and Data Analysis for Metabolomics*; Li, S., Ed.; Methods in Molecular Biology; Springer US: New York, NY, 2020; pp 265−311. DOI: 10.1007/978-1-0716-0239-3_15.

(45) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W. *Mol. Cell. Proteomics* **2011**, *10* (1), R110.000133.

(46) *HDF5*. https://www.hdfgroup.org/downloads/hdf5/ (accessed 2020-12-10).

(47) Nuñez, J. R.; Colby, S. M.; Thomas, D. G.; Tfaily, M. M.; Tolic, N.; Ulrich, E. M.; Sobus, J. R.; Metz, T. O.; Teeguarden, J. G.; Renslow, R. S. *J. Chem. Inf. Model.* **2019**, *59* (9), 4052−4060.

(48) Kyle, J. E.; Crowell, K. L.; Casey, C. P.; Fujimoto, G. M.; Kim, S.; Dautel, S. E.; Smith, R. D.; Payne, S. H.; Metz, T. O. *Bioinformatics* **2017**, *33* (11), 1744−1746.

(49) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. *Nat. Biotechnol.* **2016**, *34* (8), 828−837.

(50) Hussong, R.; Tholey, A.; Hildebrandt, A.; et al. *AIP Conf. Proc.* **2007**, *940* (1), 139−149.

(51) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825−2830.

(52) Gupta, S.; Ahadi, S.; Zhou, W.; Röst, H. *Mol. Cell. Proteomics* **2019**, *18* (4), 806−817.

(53) Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. *J. Chromatogr. A* **1998**, *805* (1), 17−35.

(54) Prince, J. T.; Marcotte, E. M. *Anal. Chem.* **2006**, *78* (17), 6140−6152.

(55) Watrous, J. D.; Henglin, M.; Claggett, B.; Lehmann, K. A.; Larson, M. G.; Cheng, S.; Jain, M. *Anal. Chem.* **2017**, *89* (3), 1399−1404.

(56) Smirnov, A.; Qiu, Y.; Jia, W.; Walker, D. I.; Jones, D. P.; Du, X. *Anal. Chem.* **2019**, *91* (14), 9069−9077.

(57) Tada, I.; Chaleckis, R.; Tsugawa, H.; Meister, I.; Zhang, P.; Lazarinis, N.; Dahlén, B.; Wheelock, C. E.; Arita, M. *Anal. Chem.* **2020**, *92* (16), 11310−11317.

(58) Yin, Y.; Wang, R.; Cai, Y.; Wang, Z.; Zhu, Z.-J. *Anal. Chem.* **2019**, *91* (18), 11897−11904.

(59) Stow, S. M.; Causon, T. J.; Zheng, X.; Kurulugama, R. T.; Mairinger, T.; May, J. C.; Rennie, E. E.; Baker, E. S.; Smith, R. D.; McLean, J. A.; Hann, S.; Fjeldsted, J. C. *Anal. Chem.* **2017**, *89* (17), 9048−9055.

(60) Lee, J.-Y.; Bilbao, A.; Conant, C. R.; Bloodsworth, K. J.; Orton, D. J.; Zhou, M.; Wilson, J. W.; Zheng, X.; Webb, I. K.; Li, A.; Hixson, K. K.; Fjeldsted, J. C.; Ibrahim, Y. M.; Payne, S. H.; Jansson, C.; Smith, R. D.; Metz, T. O. *Bioinformatics* **2021**, btab429.

(61) Blumer, M. R.; Chang, C. H.; Brayfindley, E.; Nunez, J. R.; Colby, S. M.; Renslow, R. S.; Metz, T. O. *J. Chem. Inf. Model.* **2021**, *61* (12), 5721−5725.

(62) Köster, J.; Rahmann, S. *Bioinformatics* **2012**, *28* (19), 2520−2522.