Supporting Information

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*, Thomas O. Metz*

Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA 99352 USA
* Corresponding authors

Table of Content

Terminology

Intensity values vs. dimensions vs. channels

We first define our use of "dimensions" in describing mass spectrometry data because the distinction between what qualifies as a dimension and what does not is often ambiguous. For example, in the literature, "intensity" (or "abundance") has been referred to as a dimension in addition to those dimensions readily characterized as such, e.g. *m/z*, retention time, or drift time. However, we assert that intensity/abundance represents measured signal at coordinates characterized by constituent dimensions and thus should not be considered as a distinct dimension. In other words, we define dimensionality as the number of indices needed to specify an individual element within a given array. For instance, to query LC-IMS-MS data, we must specify *m/z*, retention time, and drift time to return the value (intensity) of the corresponding array element indexed by that coordinate; thus, the data is defined by those 3 dimensions and should be considered 3-dimensional, and not 4-dimensions as has been reported in other papers.

For instance, consider a monochromatic image. Each pixel of the image has an X dimension (width), a Y dimension (height), and an intensity value (for an 8-bit image, ranging from 0 to 255). Despite each pixel being characterized by three values—X, Y, and intensity—we refer to this image as two dimensional, with the intensity values queried through X and Y indices.

Similarly, tandem MS (also known as MS/MS or MS2) has sometimes been referred to as an additional dimension. However, successive MS levels have the same indices as the accompanying MS1, and thus the same dimensionality applies. Just as MS1 can be queried by *m/z*, retention time, and drift time to yield an array element in LC-IMS-MS, so too can the MS2. We thus consider the MS2 (and, by extension, MSn) not as additional dimensions, but "channels" wherein array indices—and as it follows, array dimension—are shared. Distinction by channel then implies a parallel array structure containing the same dimensions; thus, our example of LC-IMS-MS/MS results in three-dimensional, two-channel data.

To illustrate, we return to our image analogy. In a color image, each (X, Y) coordinate contains separate values for three separate image channels: red, green, and blue. While we could refer to the color channel as a separate index and store (X, Y, channel) coordinates, we instead query values in each channel by the same (X, Y) indices and consider a color image as two-dimensional overall. This convention applies because the channel does not represent a spatial, physical, or temporal dimension that separates array elements; rather, color reflects parallel acquisitions of equal dimensionality from the camera sensor, as in MS and MS/MS.

Methods

Internal standard composition

The internal standard mixture consisted of D4-malonic acid, D4-succinic acid, D5-glycine, D4-citric acid, ^{13}C6-fructose, D5-L-tryptophan, D4-lysine, D7-alanine, D35-stearic acid, D5-benzoic acid, and D15-octanoic acid.

File input/output

For optimal use with DEIMoS, we recommend certain *msconvert* options to ensure input data replicate the vendor format as closely as possible (i.e. *msconvert.exe {filename}.{ext} -32 -z -g -outfile {filename}.mzML.gz*). Provided data in mzML format, DEIMoS parses the file contents to build a schema represented internally as a *pandas*[45,46] data frame containing arrays for each separation dimension (e.g. for LC-IMS-MS/MS: retention time, drift time, and *m/z*) and intensity.

Kernel selection

Some instruments decrease *m/z* resolution with increasing *m/z*. Even kernels of constant size can thus yield a larger *m/z* footprint. To account for such variation, DEIMoS supports scaling of other, fixed-resolution dimensions based on another dimension with dynamic resolution. Scaling is achieved by defining a reference resolution by which remaining, selected dimensions are scaled. For instance, kernel width can be scaled in the drift time dimension by *m/z* resolution.

For our example data, we determined kernel size for each dimension in two steps. First, we used a single feature of high intensity and well-defined peak shape in each dimension to define parameters for initial feature detection. The footprint of the high-intensity feature, approximately 3-sigma of a Gaussian distribution, was used to determine kernel size relative to the resolution of the underlying data. The kernel was then applied to the full dataset for rough feature coordinate extraction. Second, we sampled the features resulting from step (1) to span each dimension and produced peak statistics as a function of *m/z*. We found that peak width increases in both *m/z* and drift time dimensions with increasing *m/z*, but retention time remains largely invariant. Sampled peak statistics were used to inform final kernel size selection. See **Figure S1** for visualization of peak size analysis.

For Bruker and Waters data, a less exhaustive approach was employed in that only a handful of representative features were sampled to determine adequate feature detection parameters. For Bruker, these were 20 ppm, 0.05 V•s/cm^2, and 20 seconds for *m/z*, inverse reduced mobility, and retention time, respectively. For Waters, values were 20 ppm, 0.38 ms and 0.1 minutes for *m/z*, drift time, and retention time, respectively.

Agglomerative clustering

Agglomerative clustering is implemented via *scikit-learn* using a custom distance matrix to ensure the maximum linkage distance does not exceed the user-specified tolerance in any one dimension, i.e. Chebyshev distance. Cluster affinity is defined by complete linkage, which uses the maximum of the distances between all observations of two sets to qualify a merge. To ensure that features are merged into clusters across datasets, not within, a connectivity matrix is automatically generated to mask intra-sample linkages. However, intra-dataset clustering can occur when parent nodes unconstrained by the connectivity matrix are merged, resulting in the clustering of distal, nonadjacent child nodes. We note that nodes are not merged if the maximum linkage distance is exceeded. Thus, to prevent erroneous feature merges, the user can simply reduce their selected tolerances.

Extracted ion approach

The extracted ion approach was evaluated on deuterated internal standards with metabolites of known *m/z*. For each *m/z* of possible adducts, the data are sliced ±20 ppm to yield the extracted ion representations, and the coordinate of the most abundant feature is returned. Masses of probable adducts were calculated using the Mass Spectrometry Adduct Calculator (MSAC) for each parent molecule depending on instrument polarity: $[M+H]^+$ and $[M+Na]^+$ for positive mode, and $[M-H]^-$ for negative mode, though many more adduct types are available through MSAC.

Isotope detection

In DEIMoS, isotopes are detected in a similar fashion to alignment, wherein matching is determined within a dataset using an *m/z* offset corresponding to probable isotopic distance. Isotopic distance can be represented as the isotope mass difference (Δm) times the number of isotopic substitutions (N), divided by the formal charge (z), as shown in Equation S1.

$$\text{Equation S1:} \quad d_{iso} = \frac{N\Delta m}{z}$$

DEIMoS currently supports user supplied isotopic relationships through specification of the parameters defined in **Equation S1**, where N and z are singular values or arrays. When arrays are specified, DEIMoS enumerates isotopic distances. To evaluate the deisotoping module for singly charged analyte, we searched for ^{13}C isotopic patterns (Δm = 1.0003 Da) with up to five isotopic substitutions (N = 5) and a nominal charge of 1 (z = 1), highlighting sample-level ^{13}C isotope statistics and demonstrative features. A similar process was performed for a multiply charged analyte and an analyte with overlapping isotopic signatures, differing only in the search space of nominal charge ($z \leq 3$). In each case, per-dimension tolerances were ±20 ppm in *m/z,* ±1.5% in drift time, and ±0.3 minutes in retention time.

Partitioning

Data acquisitions of high dimensionality result in greater memory and processing demands. The feature detection process is computationally efficient for N < 3, but memory-intensive for higher-dimensional data. The data are initially stored in coordinate format, a sparse representation of an N-dimensional array, but must be converted to a dense array to support processing by convolution. To ameliorate memory limitations, partitioning functionality was implemented.

For a typical LC-IMS-MS/MS dataset, the acquired signal contains on the order of several hundred million unique coordinates on sparse representation depending on LC separation length. However, the total space is defined by 197,504 unique *m/z* values, 416 unique drift times, and 568 unique retention times, resulting in a dense array with ~5×10^{10} cells. Thus, memory requirements increase from a few gigabytes in sparse format to ~200 gigabytes in the dense representation.

To circumvent memory limitations, we slice the sparse representation by a selected dimension into partitions of user-defined size. Iteratively, each partition is cast as a dense array and processed by the algorithm of interest, such as feature detection. Partitions can be configured to overlap in order to account for edge effects of the applied convolutions. As such, the regions proximal to an artificially imposed partition edge are ignored in favor of the overlapping, non-edge regions.

The partitioning utility may also be used in applications involving feature matching, such as alignment. For example, if partitioned in 10 *m/z* increments, features from 100 - 110 *m/z* would be matched without requiring evaluation against the 120 - 130 *m/z* range, dramatically increasing computational efficiency. Additionally, as with feature detection, we account for potential errors arising from partition edge effects through partition overlap.

Plotting

Though data visualization is only a peripheral focus of this work, DEIMoS includes a plotting module with several convenience functions for common plots: stem plots for MS2 spectra, fill-between plots for 1D representations, grid plots 2D representations, and a utility to combine each into a composite plot. The composite plot visualizes, in the case of LC-IMS-MS, a 3D feature using a series of 1D and 2D plots, as shown in **Figure 2**.

Results and Discussion

Feature detection

Due to the iterative nature of lower-dimensionality approaches, computation time scales with number of features detected. For example, if N features are detected in the 2D projection of *m/z* and retention time, feature detection must be performed N times in the remaining drift time dimension. Because of resulting loop iterations, the computational load of such approaches is high despite dimensionality reduction, whereas the native dimensionality approach only requires the kernel to pass over the data once for feature detection. However, simultaneous use of all dimensions still creates the largest memory footprint, necessitating the use of a partitioning utility in cases of insufficient system memory.

Extracted ion approach

To demonstrate DEIMoS's extracted ion approach, we targeted deuterated internal standards that were spiked into the human plasma samples. Each *m/z* of interest is used to isolate features in remaining dimensions (**Figure S9)**. This operation is useful for processing data from analyses of authentic reference materials to build spectral libraries or confirm the presence of target compounds. The described approach yields a clear feature for each of the examples, confirming the presence of each internal standard based on adduct *m/z* and associated coordinates in drift and retention time. However, situations may arise in which no adducts for a given analyte are detected, or multiple features for a given adduct *m/z* appear. In such cases, further investigation is required—for example, MS2 spectra assessment—particularly in contexts involving authentic reference standards and internal standards.

Isotope detection

Isotopic signatures were identified for ^{13}C-containing analytes in a representative sample. Only signatures with at least three-member isotopologues within expected *m/z* error were considered, resulting in 2132 ^{13}C-containing adduct ions detected by their isotopic signature. Detected isotopic signatures can be then used to either (i) exclude redundant features in downstream analysis or (ii) provide further evidence supporting presence or absence of a given analyte via formula confirmation. In either case, the ability to rapidly identify these signatures is critical to mass spectrometry analysis. Representative examples for singly charged, multiply charged, and overlapping isotopic signature are depicted in **Figures S10**, **S11**, and **S12**, respectively.

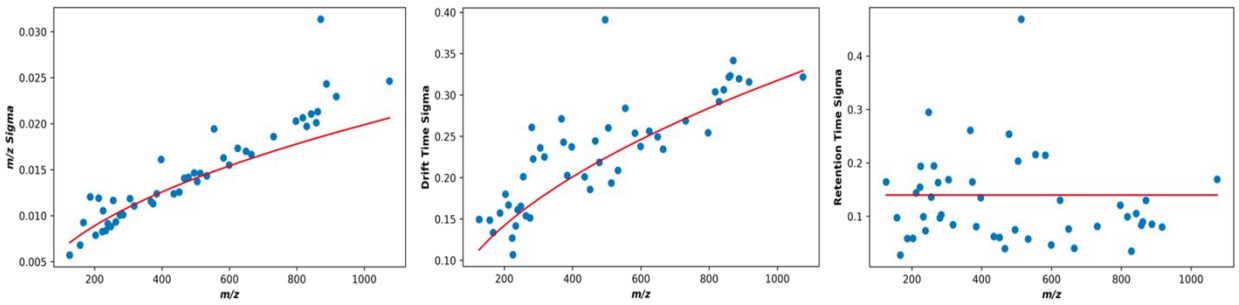

Figure S1. Peak width characterization. Features selected to span each separation dimension to determine a relationship between *m/z* and peak width. The sigma of the distributions in *m/z* and drift time (blue) are plotted alongside the relative resolution of the sampling interval in *m/z* (red), illustrating the relationship between dynamic resolution and dynamic kernel size. For retention time (blue), as there is no discernable relationship between separation dimensions and peak width, the average width of the sampled distributions (red) was used to inform kernel size selection.

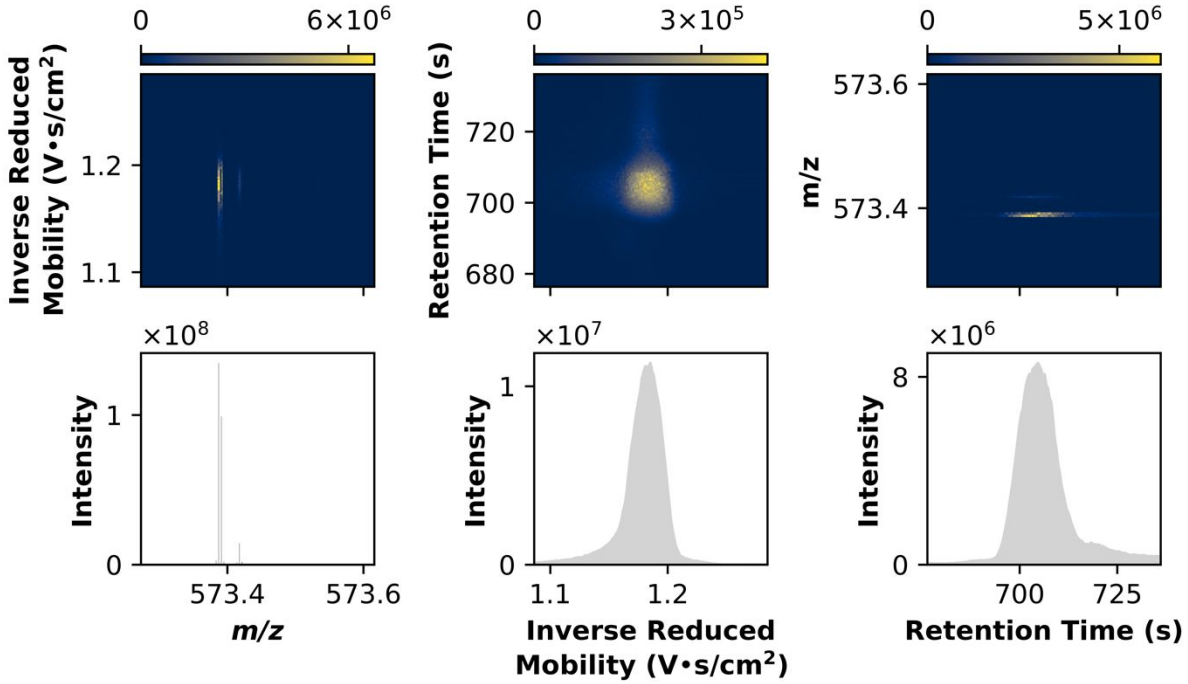

Figure S2. Representative feature, Bruker. An example feature from data produced by a Bruker timsTOF Pro, visualized in the native dimensionality of LC-IMS-MS. The top row shows 2D representations of the data, left to right: *m/z* versus drift time, drift time versus retention time, and retention time versus *m/z.* The bottom row shows 1D representations, left to right: *m/z*, drift time, retention time. Each panel is the result of summing across dimensions not shown.

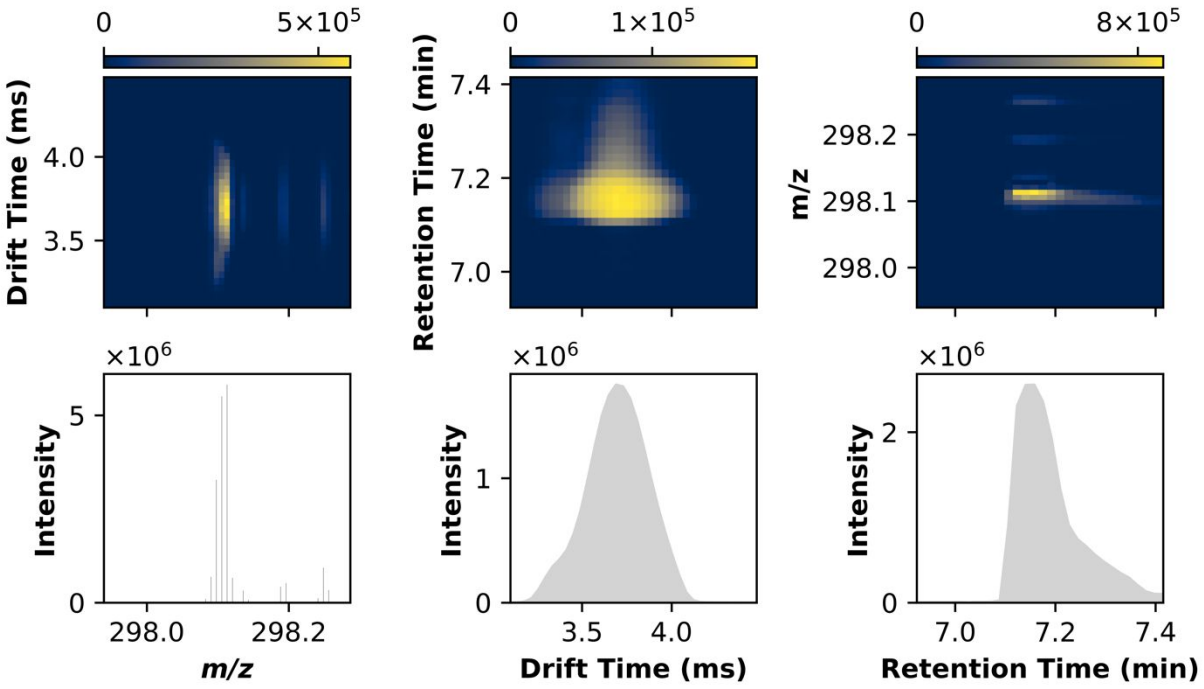

Figure S3. Representative feature, Waters. An example feature from data produced by a Waters UPLC i-Class coupled to a Synapt G2Si, visualized in the native dimensionality of LC-IMS-MS. The top row shows 2D representations of the data, left to right: *m/z* versus drift time, drift time versus retention time, and retention time versus *m/z*. The bottom row shows 1D representations, left to right: *m/z*, drift time, retention time. Each panel is the result of summing across dimensions not shown.

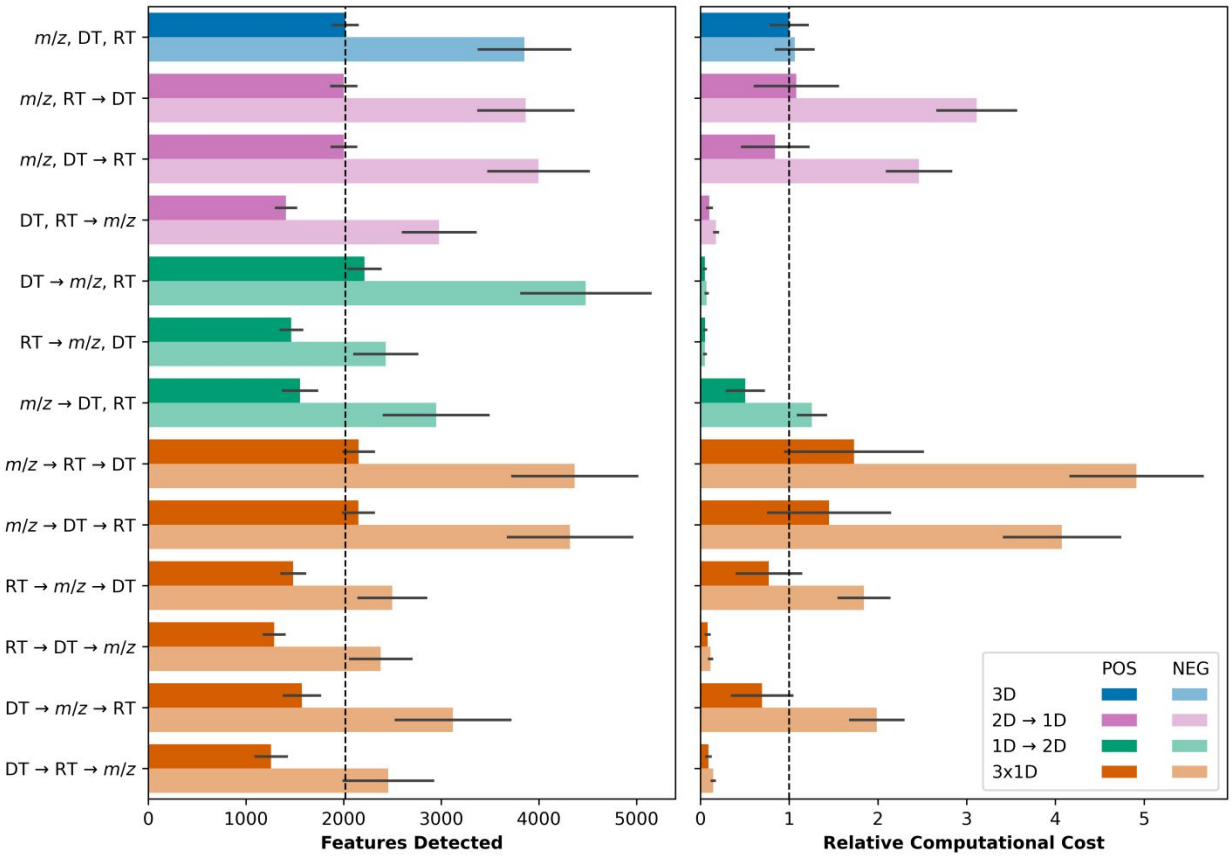

Figure S4. Comparison of peak detection in all dimension projections. Evaluation of number of features detected (left) and relative computational cost (right) in native dimensionality (for LC-IMS-MS, 3D, blue), as well as all possible lower dimensional projections. Permutations of (i) 2D followed by 1D (pink), (ii) 1D followed by 2D (green), and (iii) iterative 1D approaches (orange) are shown for positive (dark) and negative (light) mode, respectively. Feature counts and computational costs were averaged over all acquired samples, per ionization mode, with one standard deviation represented by the error bars. Computational cost was normalized relative to 3D, positive mode.

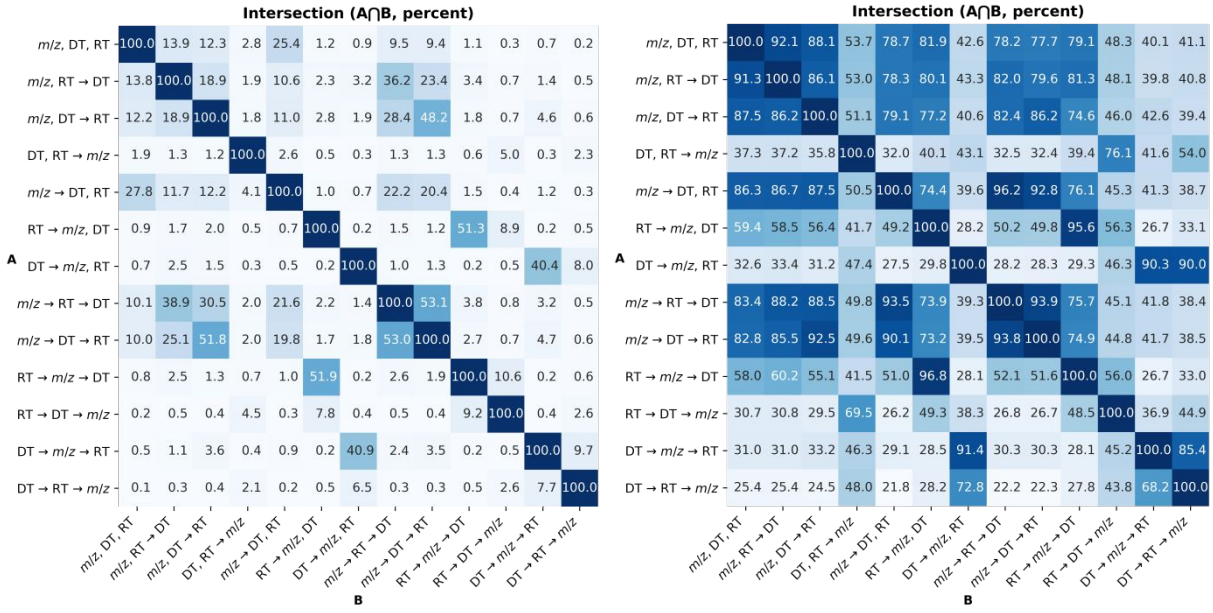

Figure S5. Comparison of peak detection similarity. Each peak detection approach for positive ionization mode was compared pairwise to assess intersection, or peaks shared between methods, averaged across replicates and samples (N=168). In the left panel, peak coordinates were required to match to instrument precision, resulting in few intersecting peaks among methods. In the right panel, we employed a match tolerance of ±20 ppm, ±1.5%, and ±0.3 minutes for *m/z*, drift time, and retention time, respectively, based on tolerances used for cross-sample alignment. Negative ionization mode results did not materially differ.

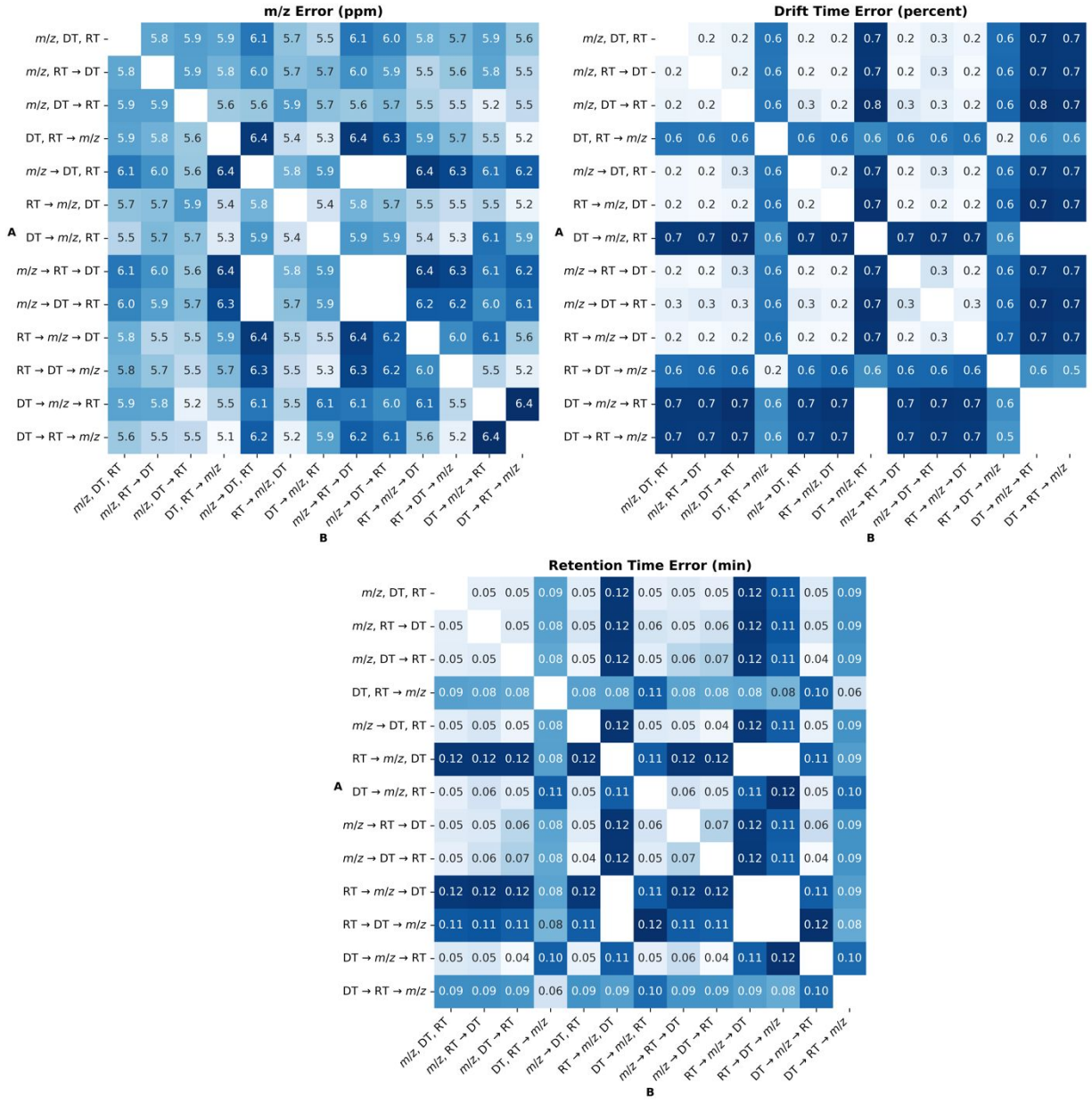

Figure S6. Comparison of peak coordinate deviation. Each peak detection approach for positive ionization mode was compared pairwise to assess the difference in corresponding peak apices, averaged across replicates and samples (N=168). Coordinates that did not differ were excluded from the average. Top right, top left, and bottom panel describe such error for each dimension: *m/z* in parts per million, drift time in percent, and retention time in minutes. Negative ionization mode results did not materially differ.

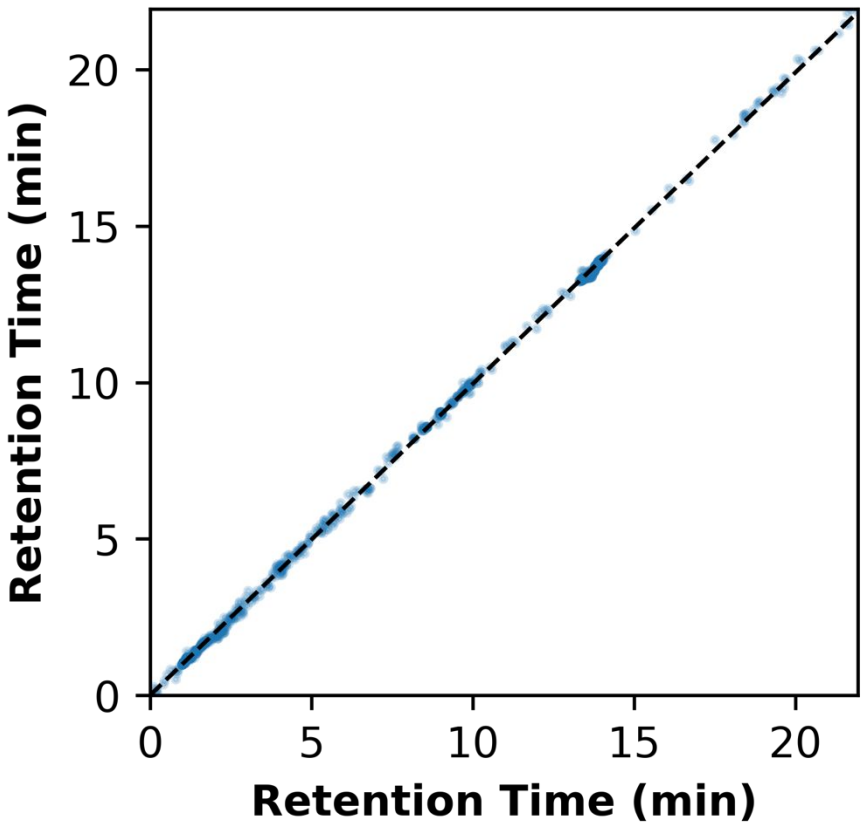

Figure S7. Linear alignment by support vector regression. Support vector regression (SVR) was evaluated here on the retention time dimension between 2 illustrative samples described by a linear relationship in retention time. Accordingly, to model this relationship, a linear kernel was selected.

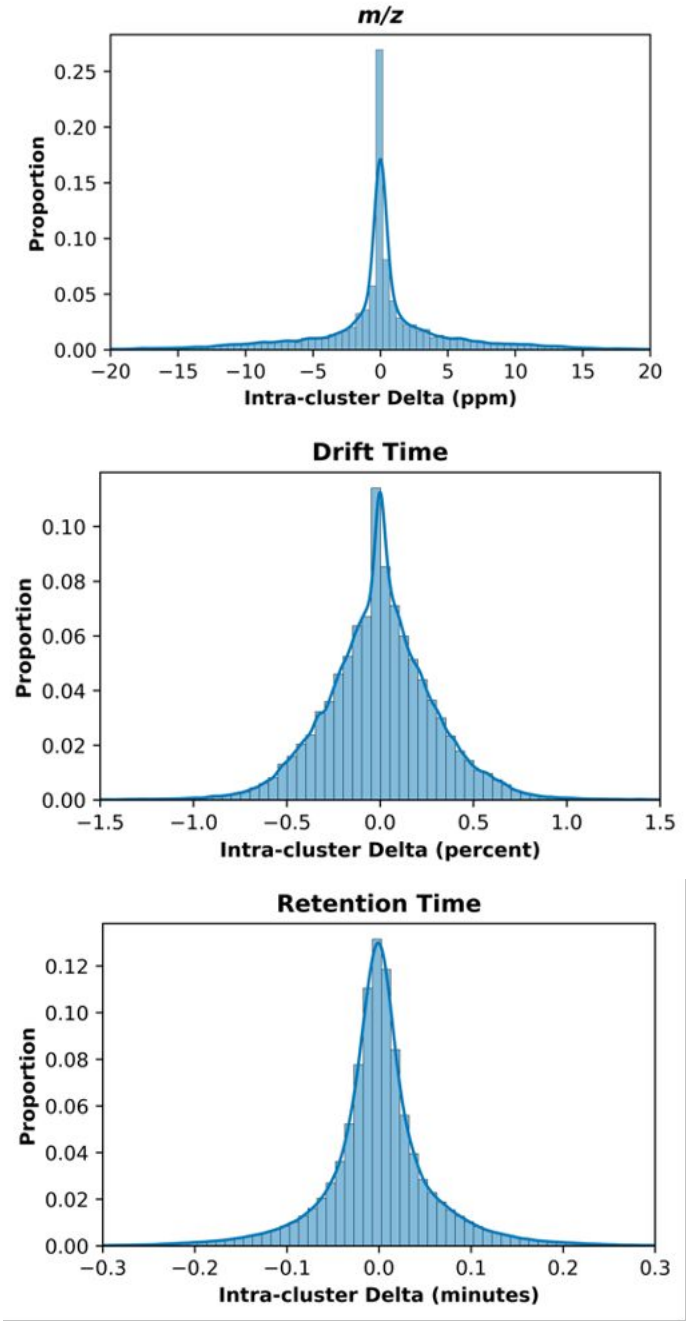

Figure S8. Alignment characterization. Features defined in three dimensions after alignment by agglomerative clustering were assessed for intra-cluster variance. That is, the average error within a feature cluster following comparison to the mean of the cluster in that dimension. The above shows the histogram and kernel density estimation of said error in each dimension, reported as ppm for *m/z*, percent for drift time, and minutes for retention time.

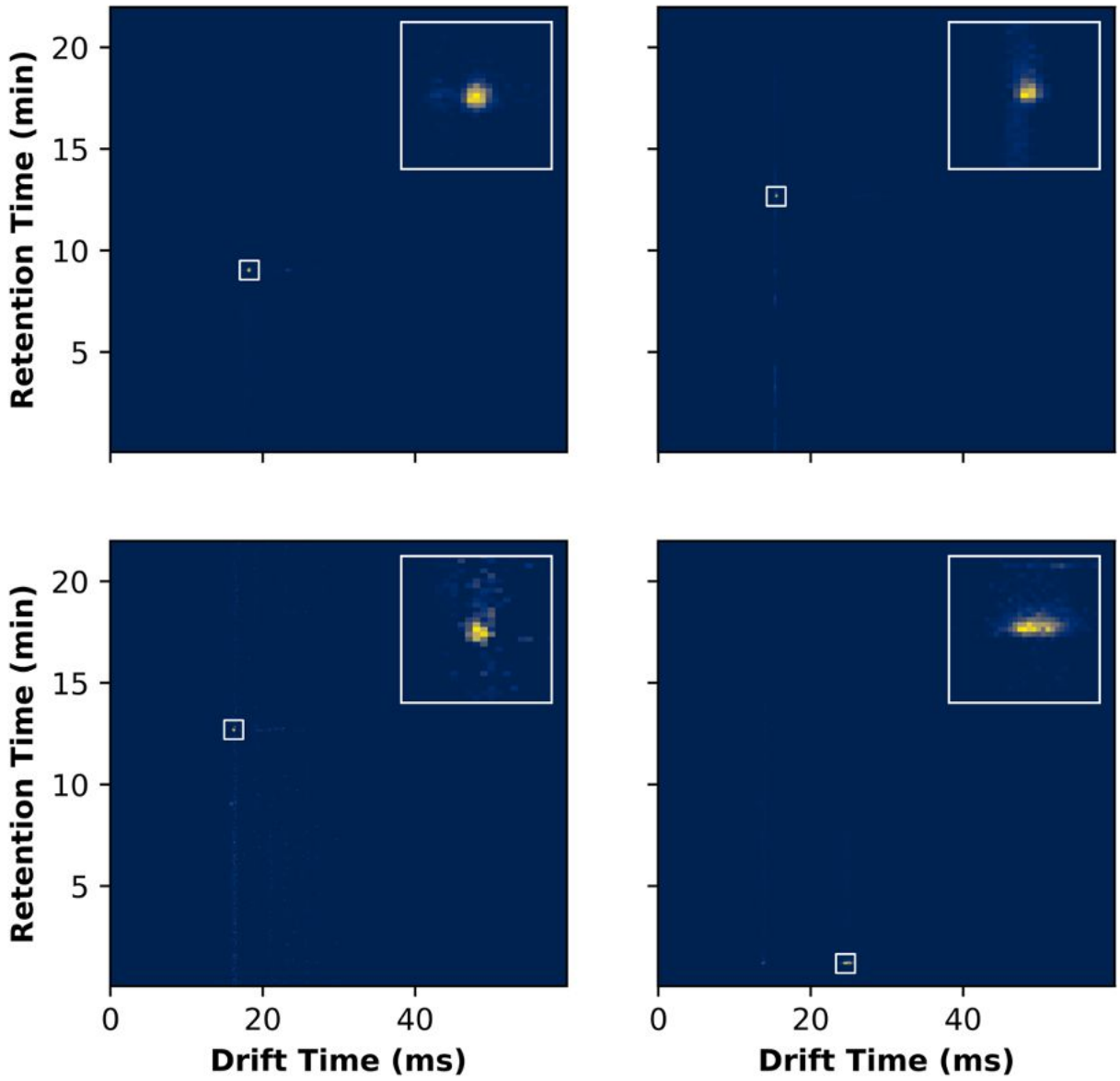

Figure S9. Extracted ion approach. Applying DEIMoS's suite of targeted functionality, extracted ion representations of the data are easily generated for presumably present features wherein one or more separation coordinates are known. In this case, deuterated internal standards with known *m/z* were analyzed using LC-IMS-MS/MS, leaving corresponding drift and retention times as unknown coordinates. By isolating the data in *m/z* for probable adduct masses (±20 ppm tolerance), present analytes will appear in the resulting 2D representations. Here we visualize the deuterated forms of L-tryptophan $[M+H]^+$ (210.1285 *m/z*, top left), lysine $[M+H]^+$ (151.1379 *m/z*, top right), lysine $[M+Na]^+$ (173.1198 *m/z*, bottom left), and alanine $[M+Na]^+$ (119.0808 *m/z*, bottom right), each indicated by white highlight and zoomed for emphasis in the inset panel.

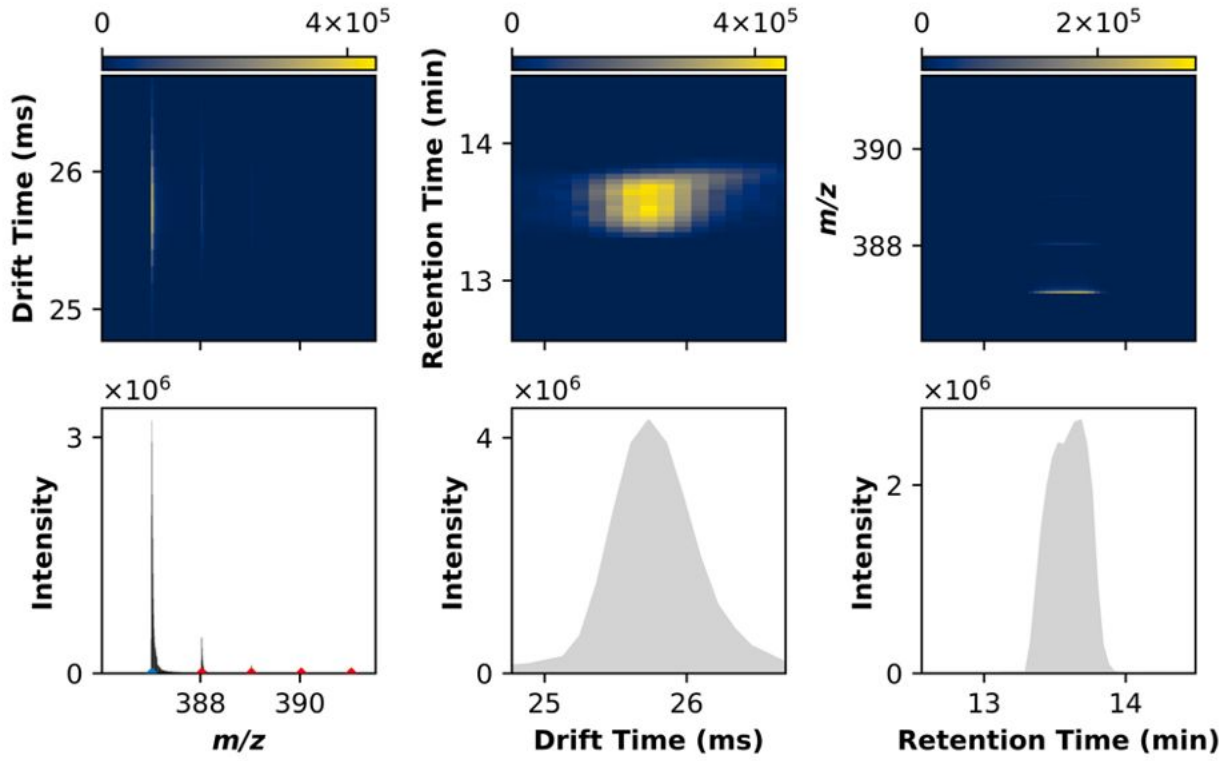

Figure S10. Isotope detection, singly charged. An example of isotope detection in the native dimensionality of LC-IMS-MS data. The top row shows 2D representations of the data, left to right: *m/z* versus drift time, drift time versus retention time, and retention time versus *m/z*. The bottom row shows 1D representations, left to right: *m/z*, drift time, retention time. Each panel is the result of summing across dimensions not shown. In the 1D plot of *m/z*, the parent ion is indicated in blue, with isotopologues highlighted in red. By this representation, it is apparent that isotopologues, for the resolution of this LC-IMS-MS/MS experiment, are indistinguishable in drift and retention time.

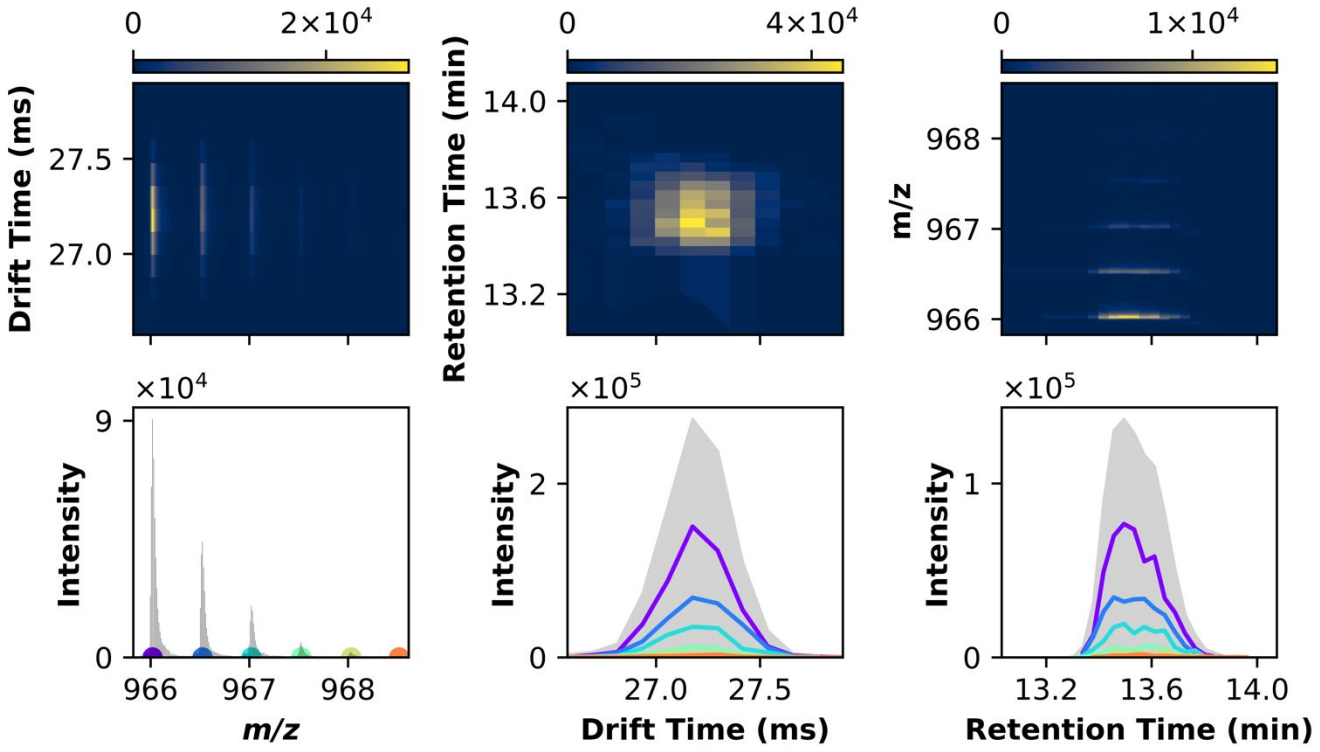

Figure S11. Isotope detection, multiply charged. An example of isotope detection of a multiply charged analyte (z = +2) in the native dimensionality of LC-IMS-MS data. The top row shows 2D representations of the data, left to right: *m/z* versus drift time, drift time versus retention time, and retention time versus *m/z*. The bottom row shows 1D representations, left to right: *m/z*, drift time, retention time. Each panel is the result of summing across dimensions not shown. In the 1D plot of *m/z*, the parent ion is indicated in purple, with isotopologues highlighted in according to a rainbow spectrum (left to right: blue, cyan, green, yellow, orange). Profiles of each colored isotopologue correspond in the 1D projections of drift and retention time. Again, these isotopologues, for the resolution of this LC-IMS-MS/MS experiment, are indistinguishable in drift and retention time.

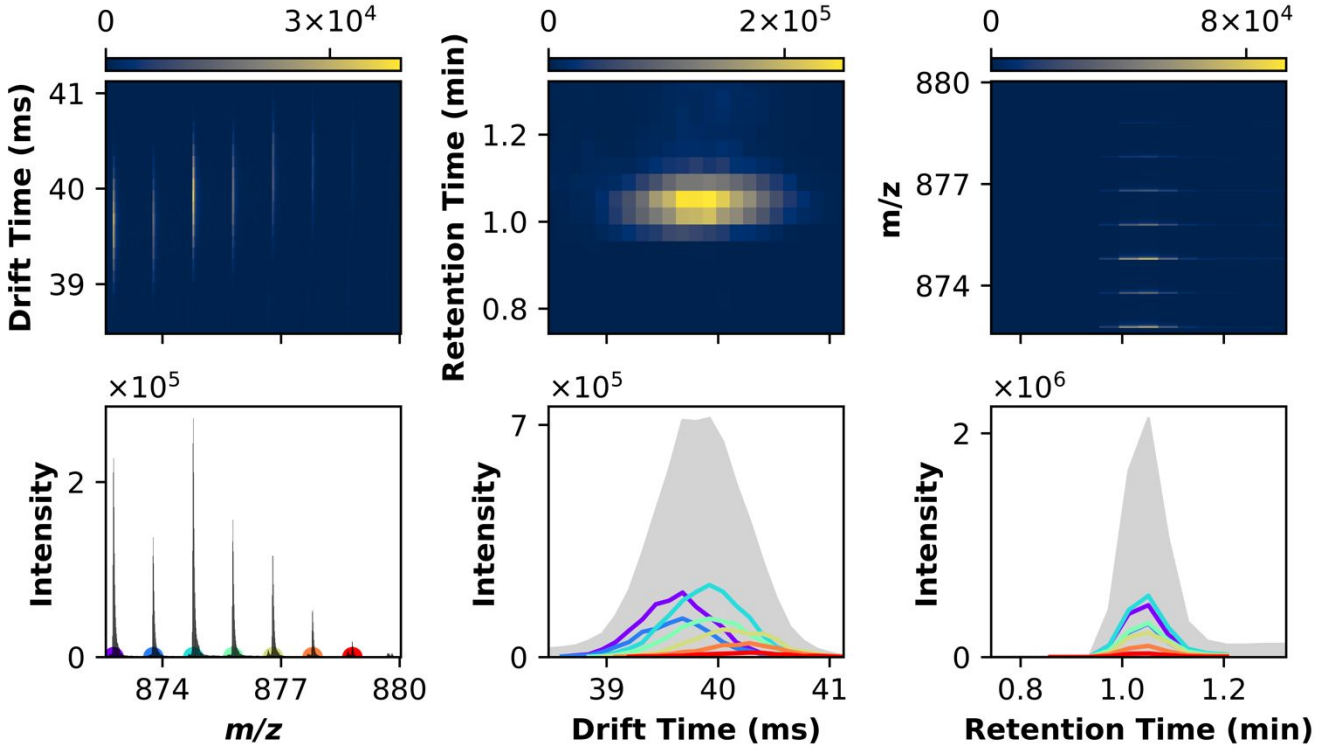

Figure S12. Isotope detection, overlapping. An example of isotope detection overlapping, singly charged analytes (z = +1) in the native dimensionality of LC-IMS-MS data. The top row shows 2D representations of the data, left to right: *m/z* versus drift time, drift time versus retention time, and retention time versus *m/z.* The bottom row shows 1D representations, left to right: *m/z*, drift time, retention time. Each panel is the result of summing across dimensions not shown. In the 1D plot of *m/z*, the parent ions are indicated in purple and cyan, respectively, with isotopologues highlighted in according to a rainbow spectrum (left to right: blue, cyan, green, yellow, orange, red). Note the cyan feature is considered both an isotopologue of the purple feature, as well as parent to isotopologues to its right. Profiles of each colored isotopologue correspond in the 1D projections of drift and retention time. Here, the overlapping analytes were not distinguishable by retention time, but were separable by drift time, as there were 3 distinct drift time groupings. These represent either 3 parent analytes, each separated by its drift time profile grouping, or two parent analytes, wherein the middle drift time groupings (cyan, green) represent a linear combination of co-drifting populations (blue and purple, and yellow, orange, and red, respectively).