2020

Generating Peptide Mass Spectrometry Ground Truth Data

Jessica L. Henning
The University Of Montana

Rob Smith
The University Of Montana

GENERATING PEPTIDE MASS SPECTROMETRY GROUND TRUTH DATA

By

Jessica Lee Henning

Bachelor of Science, University of Montana, Missoula, MT, 2017

Thesis

presented in partial fulfillment of the requirements
for the degree of

Master of Science
in Computer Science

The University of Montana
Missoula, MT

May 2020

Approved by:

Scott Whittenburg, Dean of The Graduate School
Graduate School

Rob Smith
Computer Science

Michael Cassens
Media Arts

Doug Brinkerhoff
Computer Science

Table of Contents

Henning, Jessica, M.S., Spring 2020 Computer Science

Abstract - Generating Peptide Mass Spectrometry Ground Truth Data

Chairperson: Rob Smith

Mass spectrometry (MS) uses mass-to-charge ratios of measured particles to decode the identities and quantities of molecules in a sample. Interpretations of raw MS depends upon data processing software that renders it human-interpretable. Quantitative MS workflows are complex experimental chains and it is crucial to know the performance and bias of each data processing method as they impact accuracy, coverage, and statistical significance of the result.

Although existing MS workflows are ubiquitous, many practitioners encounter data processing results that question current workflow accuracy. Benchmark datasets are often used to quantitatively assess the strengths and weaknesses of existing solutions and to create and develop new solutions.

Quantitative evaluations of data processing are scarce, in part because of the scarcity of ground truth data available. The properties of MS data make ground truth especially difficult to generate. These challenges can be grouped into human-related challenges and computational challenges. Obtaining ground truth requires tools that mitigate these challenges, allowing users to quickly and accurately manually annotate data.

For this project, we present JS-MS 2.0, a software suite that provides a dependency-free, browser-based, one click, cross-platform solution for creating precursor ground truth. The software retains a previous version's capacity for loading, viewing, and navigating MS1 data in 2- and 3-D, and adds tools for capturing, editing, saving, and viewing isotopic envelope and extracted isotopic chromatogram features.

We also provide a novel ground truth dataset for mass spectrometry data analysis at the precursor (MS1) signal level comprised of isolated peptide signals from UPS2, a popular complex standard for proteomics analysis, requiring more than 1,000 hours of manual curation. The dataset consists of more than 62 million points, with 1,294,008 grouped into 57,518 extracted ion chromatograms, and those grouped into 14,111 isotopic envelopes using JS-MS 2.0.

This dataset can be used to quantify many evaluations such as extracted ion chromatograms (XIC) extraction algorithms, XIC clustering into isotopic envelopes, MS1-based quantification methods, MS2 quantification methods, and false detection estimations. JS-MS 2.0 will allow for the creation and validation of more ground truth datasets that will assist further evaluation and algorithm creation for mass spectrometry data.

A Web-Based System for Creating, Viewing, and Editing Precursor Mass Spectrometry Ground Truth Data

Jessica Henning and Rob Smith

Abstract

Very few quantitative evaluations exist for precursor mass spectrometry data due to the lack of tools for enabling the manual feature finding necessary to generate this data. Other lacks the ability to capture, edit, save, and view precursor mass spectrometry data. We present JS-MS 2.0, a software suite that provides a dependency-free, browser-based, one click, cross-platform solution for creating precursor ground truth. The software retains the first version's capacity for loading, viewing, and navigating MS1 data in 2- and 3-D, and adds tools for capturing, editing, saving and viewing isotopic envelope and extracted isotopic chromatogram features. The software can also be used to view and explore the results of feature finding algorithms. JS-MS 2.0 enables faster creation and inspection of precursor mass spectrometry ground truth data. It is publicly available with a GPL 2.0 license at github.com/optimusmoose/jsms.

Background

Mass spectrometry (MS) is a powerful vector for the analysis of molecular components (such as proteins, peptides, lipids, and metabolites) in biological samples across a broad range of applications [1]. MS experiments generate datasets consisting of millions of 3-D points consisting of mass-to-charge (m/z), retention time (RT), and intensity. MS experiments require the mapping of all or some of these points to signal groups that correspond to a single (or multiple, in the case of isomers) molecules at a given charge state.

This process, called feature detection, has been addressed by numerous algorithms, commercial software, and public software such as MaxQuant [2], MZMine 2 [3], CentWave (XCMS) [4], MatchedFilter (XCMS) [4], and Massifquant (XCMS) [5]. Unfortunately, many of these and other algorithms for MS1-aware analysis have not been quantitatively evaluated [6], mostly due to the fact that ground truth data is very difficult to generate.

A system for producing precursor ground truth annotations requires several functions:

- It must parse, load, store, and retrieve precursor data.
- It must efficiently display many points on the screen.
- It must display points in representative subsets, as not all points can be rendered on the screen at once.
- It must output the data in easy-to-port formats.
- It must provide the user with efficient navigation of the data (zoom and shifting to the right, left, up, or down).

It should also be designed in such a way as to allow easy cross-platform install without the need for excessive dependencies or onerous compilation.

To date, the only such system is JS-MS [7]. JS-MS is a browser-based JavaScript viewer and Java server designed to load and view precursor mass spectrometry data. It provides useful navigation tools such as the ability to zoom in, zoom out, pan up-down-left-right, and toggle between 2-D and 3-D views.

The server component communicates with the view through a simple JSON API, which makes it interchangeable with any other server that implements the same API. The server responds to queries for specific (m/z, RT) windows. Each query includes a requested limit on the number of points returned, which invokes the server's algorithm for selecting a representative subset of points, allowing for the user to view the characteristics of the data while only seeing a portion of the points in the given (m/z, RT) region. The server implements the MzTree data structure [8], which is a modified R-Tree that organizes the MS1 points in alternating sorting of m/z and RT to provide fast query response whether the data region requested is primarily across m/z, RT, or both.

JS-MS is packaged as a single self-contained JAR, and the only dependencies are the Java Runtime Environment (JRE) and a web browser, both typically already available on any computer.

Since the publication of JS-MS, our group has substantially extended the software. In addition to loading, viewing, and navigating MS1 data in 2-D and 3-D, JS-MS 2.0 extends JS-MS by providing tools for creating, editing, and viewing annotations of extracted ion chromatograms (called isotopic traces hereafter) and features (called isotopic envelopes hereafter). These tools facilitate inspection and modification of algorithms for isotope trace and isotopic envelope annotation, as well as the creation of manually annotated precursor ground truth.

To date, our group has used JS-MS 2.0 to create the first ever quantitative ground truth dataset for MS1 data [9], as well as the first quantitative evaluation of algorithms that group traces into isotopic envelopes [10]. We are now releasing JS-MS 2.0 in hopes that others will use it to generate more ground truth to enable new and more extensive quantitative evaluations of MS1 algorithms.

Implementation

JS-MS 2.0 extends the original JS-MS implementation through many extensions to the view, additions to the MzTree data storage and retrieval system, and additional API calls to the server.

The JS-MS 2.0 view provides extensions that enable annotations to be displayed, recorded, and edited, as well as helper tools that facilitate fast annotation decisions and annotation inspection. The original application included logic that colorized signals based on intensity. In JS-MS 2.0, additional logic defines color based on isotopic trace or isotopic envelope membership (in each mode, respectively) such that proximate signals have different colors.

JS-MS 2.0 has a new ruler feature written in Three.js [11] that calculates the expected m/z intervals of an envelope of a given charge state. Vertical lines are drawn where each trace should appear with m/z gaps adjusting according to the charge state indicated by the user. Bounds checking prevents the ruler from extending beyond the plot range. Ratios are calculated to enable scaling of the ruler on zoom. On-click events activate the ruler when any number key is pressed and deviated when the tilde key is pressed.

The bookmark list is a new feature implemented in JavaScript and HTML that provides a means of storing, editing and applying (m/z, RT) coordinates to facilitate fast navigation to regions of interest. JavaScript calls dynamically add and remove rows from the table, edit entries in the table, and store the table in the cache. A parser function inputs a tab separated text file of bookmarks. An export function writes out the current bookmark list in the same tab separated format.

Annotations of isotopic traces use a rectangle tool that is written in Three.js. On-click and draw JavaScript functions map the region of data traced by the user's mouse to (m/z, RT) coordinates used to update the point membership. An on-click event for the control key is used to toggle the function of the rectangle to remove points from an isotopic trace.

Annotations of isotopic envelopes require the selection of one or more isotopic traces to group into an isotopic envelope. Because users will not likely click directly on a point in a trace, trace selection relies on a JavaScript function that finds the closest trace within a threshold to the point clicked. Alternatively, users can click and drag a line through multiple traces to perform the same process across a set of points. An on-click event for the control key is used to toggle the behavior of the on-click event to remove one or more traces from an envelope.

Since traces tend to occur in straight lines along a given m/z, guidelines can be drawn using the Guard Rails feature. Using Three.js parallel lines are drawn along the m/z for a given m/z width with the appropriate projection in 2-d or 3-d mode, allowing the feature to persist independent of graph panning, rotation, or zoom. On-click events activate the tool with the 'g' key and deactivate with the 'h' key.

JS-MS 2.0 also includes extra controls for the user to modify view parameters such as point threshold, logarithmic height scaling, and label precision. The point threshold is a function implemented in Java that limits the number of points rendered in a given view, selecting a representative subset of points using the weighted striding algorithm [8]. Applying a point threshold allows for faster load time and graph navigation. Logarithmic height scaling is implemented with JavaScript and mathematical functions that scale point intensity by a logarithmic factor to facilitate greater contrast between signal and background noise. Label precision is also implemented with JavaScript to decrease or increase the level of precision for (m/z, RT) coordinates. This function rounds the (m/z, RT) coordinates to the desired accuracy from the user.

The MzTree data structure is a modified R-Tree [8] that interleaves data partitions sorted by RT and m/z for fast queries in either dimension. The previously published version of the data structure did not include the fields required for annotation (such as isotope trace ID and isotopic envelope ID). The previous version also lacked a new index of points sorted by intensity which is used in the jump button (discussed later)

The original JS-MS featured an HTTP API that included functions to retrieve a subset of points given an (m/z, RT) window, with an optional limit on the number of points returned. The API was extended to include isotopic trace and isotopic envelope annotation fields in the returned JSON data as well as functions to assign and edit those fields.

Results

The user interacts with JS-MS through three main interfaces: the graph interface, the control panel, and the parameter panel.

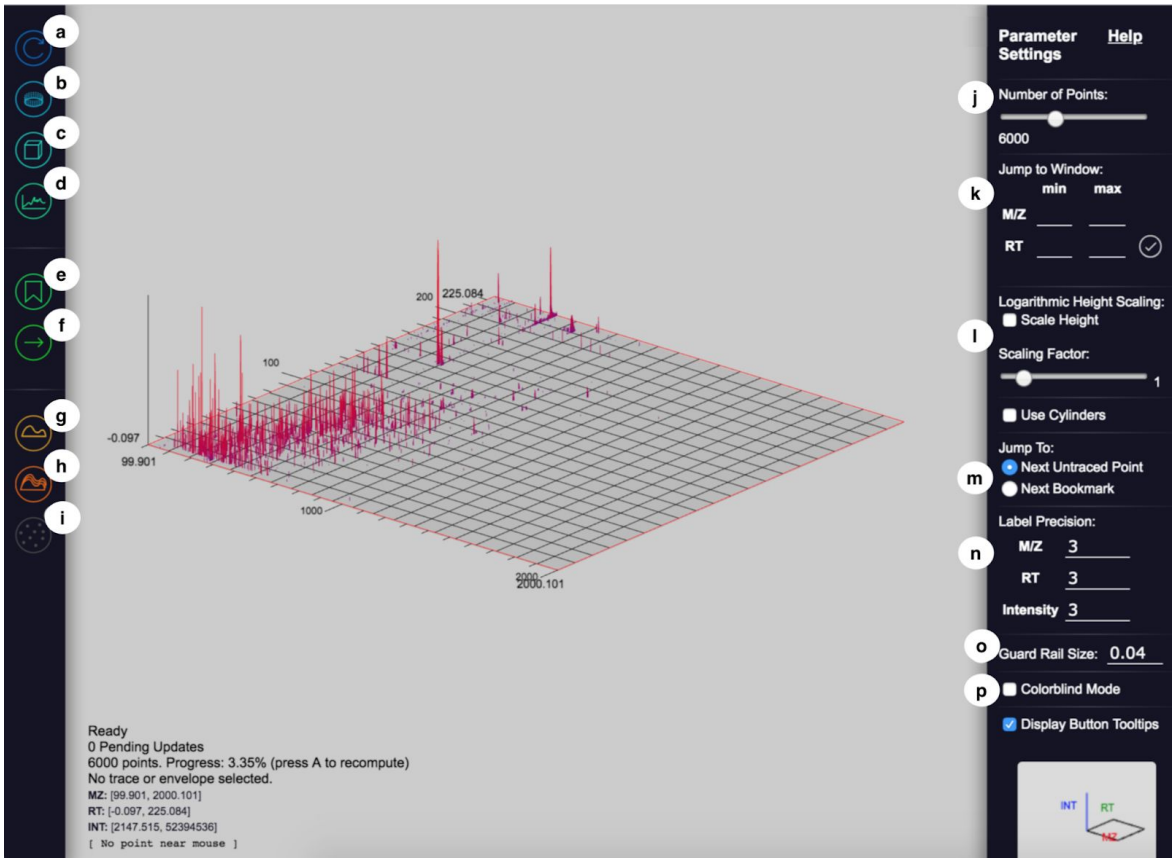


Figure 1: JS-MS 2.0 interface includes a control panel (left), graph interface (center), and parameter panel (right).

Graph Interface

The principal purpose of the graph interface is twofold: First, it displays mass spectrometry points and isotopic trace and isotopic envelope annotations of these points. Second, it provides the controls for recording and editing these annotations.
Users have the ability to navigate to areas of interest through several means. First, the user can pan, zoom in and out, and toggle between 2-D and 3-D views of areas of their choice.

Second, users can build a bookmark list to enumerate data points of interest by listing (m/z, RT) coordinates or using the "select current location" button (see Figure 2e), which will add the current location to a list of one-click navigable data regions called the jump list. The jump list provides a useful mechanism for rapidly navigating to areas of interest. For example, if third party software provides a list of regions with poor feature detection, low intensity features, or known compounds, the jump list can be used to quickly iterate through the inspection of each corresponding data region. The bookmark interface features a button for importing and exporting bookmark lists in .tsv format (see Figure 2a-b), and each bookmark entry can be edited or deleted (see Figure 2c-d).

Third, users can use the jump button to navigate to other data areas. There are two functions associated with the jump button that can be toggled in the parameter panel. The first jump function is to jump to the next highest intensity point that is not part of an annotation. By clicking it, the graph will respond by displaying the area around the point, which will be denoted by an X on the graph. The second jump function is used to enumerate through the graphs of the areas around the points listed in the bookmark list. Using this feature, users can quickly inspect many envelopes or other data features in which they may be interested.

Fourth, a convenient "jump to window" mechanism on the control panel allows users to specify the exact window that they would like to display. This functionality facilitates the creation of reproducible graphs for inspection and publication.

Control Panel

The control panel contains interfaces for users to define the graph's behavior to match their intended purpose.

- Refresh (see Figure 1a). This button reloads the data on the view from the server.
- View all data (see Figure 1b). This button displays the entire data set.
- Toggle 2-D/3-D (see Figure 1c). This button switches the graph display mode and redraws the graph.
- Ion current view (see Figure 1d). This button rotates the view to a 2-D projection of the 3-D view such that the x-axis is m/z and the y-axis is intensity.
- Bookmarks (see Figure 1e). This button shows or hides the bookmark interface.
- Jump (see Figure 1f). This button's functions are described above in "View."
- Trace mode (see Figure 1g). This button activates annotation mode, which is described below.

- Envelope mode (see Figure 1h). This button activates isotopic envelope annotation mode, which is described below.
- Mark as noise button (see Figure 1i). This button is used to indicate that all distinguishable signals in a view have been annotated and is further described below.

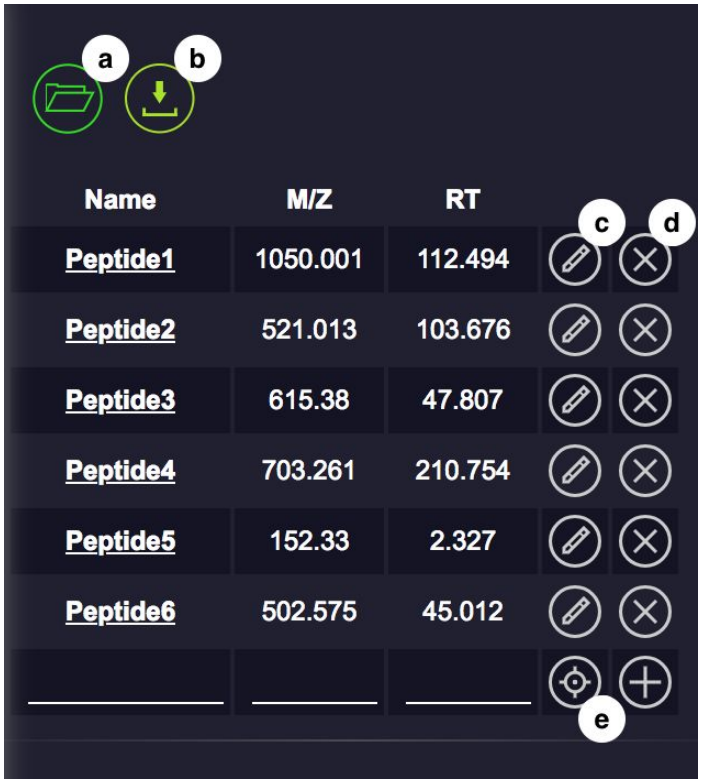


Figure 2: JS-MS 2.0 Bookmark List allows users to easily create, navigate to, import, and export a list of (m/z, RT) coordinates.

Parameter Panel

The parameter panel contains settings that adjust the view.

- Point threshold (see Figure 1j). Users can control how many points are rendered for the given view. In the event the setting is lower than the actual number of points, JS-MS selects a representative subset of points using the weighted striding algorithm described in [8].
- Set view window (see Figure 1k). This tool allows users to obtain a consistent view given the same specified (m/z, RT) window.
- Height scaling (see Figure 1l). The height scaling slider changes the intensity and colorization scaling of points in order to more effectively display low intensity points in 3-D mode.
- Jump options (see Figure 1m). This button specifies which jump function is active.
- Label precision (see Figure 1n). This setting controls how many digits of precision are displayed on the graph.

- Guard rail size (button see Figure 1o, use case see Figure 3b). The guard rail is a set of parallel lines that can be displayed for a given m/z to assist in annotating an isotopic trace. This setting controls the width of the lines.
- Colorblind mode (see Figure 1p). The colors used by the system can be limited to those visible to colorblind people.

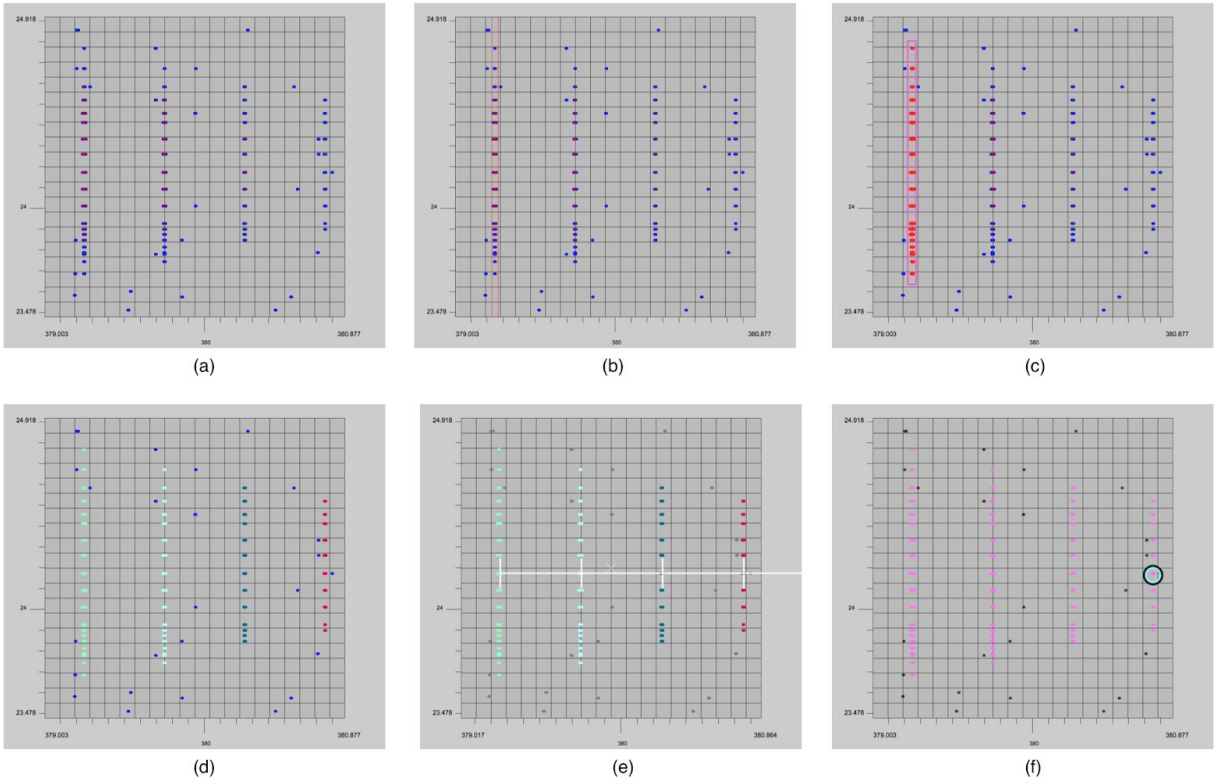


Figure 3: The annotation process for isotopic traces and envelopes. (a) Signals are shown in color based on intensity. (b) Guard rails are used to help distinguish which signals belong in an isotopic trace. (c-d) In trace mode, users mark which signals belong to an isotopic trace. (e) The ruler shows a user specified charge state to measure the isotopic traces within an isotopic envelope. (f) In envelope mode, users can specify all the isotopic traces belonging to an isotopic envelope.

Annotation

Isotopic Trace Mode. When a user enters isotopic trace mode, they are given the option to create a new trace or select an existing trace to edit. Each time a new trace is created, the trace is given an ID and color. Users select the points belonging to the trace by clicking and dragging a rectangle over the desired points to highlight them in the given color (see Figure 3c). The same procedure is used to edit an existing trace, only the control key is depressed while drawing the rectangle.

Isotopic Envelope Mode. After the user has identified isotopic traces, they can group them together with isotopic envelope mode. Similar to isotopic trace mode, this mode creates a new

envelope ID and color for each new envelope created. The user then selects all isotopic traces that belong to the same group (see Figure 3f). Isotopic traces can be grouped by clicking each trace or simply by dragging a line across all traces in an envelope. To help the user distinguish which isotopic traces belong together, the ruler tool shows m/z intervals corresponding to specific charge states. The ruler will appear wherever the mouse is placed when users select a number from the keypad. The ruler moves with the graph as the user zooms or pans, and will remain present until the user hits the tilde key. The m/z distance displayed is 1/z, where z is the number selected and the charge state of a hypothetical compound at the given mass (see Figure 3e). Users can also toggle between 2-D and 3-D mode while in either isotopic trace or isotopic envelope mode to ensure peak alignment. Isotopic traces can be added to existing isotopic envelopes at any time following this procedure and they can be removed in the same way while depressing the control key.

Mark as Noise Button. When all distinguishable points in the current view have been annotated the user can mark all other points in the view as noise. When a point is marked as noise it will be colored gray in the view and given an ID of -1 when exported to .csv. To prevent users from marking unseen points as noise, the graph view must be displaying a number of points below the point threshold to ensure that the user is viewing every point within the (m/z, RT) coordinates and none are hidden.

Discussion

Algorithm performance can significantly affect the results of mass spectrometry experiments [12], and as such, a performance evaluation should be part of any new algorithm publication. The current workflow for algorithm evaluation typically reports performance based on consensus results [13]. While consensus results provide a qualitative gauge of how similar result sets are, they do not answer the critical question--how accurate are these results? Instead, consensus results measure how closely new algorithms perform compared to prior ones. While a positive consensus result does measure similarity to previous performance, it can't distinguish whether differences are due to improvement or decline in accuracy.

The creation of benchmark datasets for precursor-aware mass spectrometry algorithms with JS-MS 2.0 will enable a new workflow for precursor MS algorithm evaluation that includes quantitative evaluation. New algorithms can be designed using information derived from ground truth annotations created with JS-MS. Once implemented, their performance can be evaluated in terms of, for instance, m/z accuracy of traces annotated, to demonstrate clear improvement over existing algorithms. These evaluations will demonstrate strengths and weaknesses to reviewers and users alike.

One such benchmark dataset is currently being constructed by our group for isotopic trace algorithms, and the community is invited to use JS-MS 2.0 to create many more such datasets for any and all precursor-aware applications (such as quantification, centroiding, etc.).

Conclusion

JS-MS 2.0 provides a dependency-free, browser-based, cross-platform solution for creating MS precursor ground truth. Novel interfaces allows users to quickly add, edit, import, and export isotope trace and isotopic envelope annotations.

While other MS viewers do not allow users to easily navigate MS datasets, the innovative navigation tools in JS-MS 2.0 give users the ability to inspect and annotate any area of signals quickly with pan, zoom, and bookmark lists. It combines interactive 2-D and 3-D plots with fast, easy to use navigation tools allowing for manual annotation of even the largest MS datasets. The creation of ground truth MS datasets will benefit algorithm development, quantitative evaluation, and help practitioners assess strengths and weaknesses of existing workflows.

JS-MS 2.0 is implemented as a JavaScript front-end and Java back-end, it is lightweight with browser-based cross-platform compatibility.

References

1. Cole, R.B.: Electrospray Ionization Mass Spectrometry: Fundamentals, Instrumentation, and Applications. Wiley, N.p. (1997)

2. Cox, J., Mann, M.: MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology **26**(12), 1367–1372 (2008)

3. Pluskal, T., Castillo, S., Villar-Briones, A., Oreˇsiˇc, M.: Mzmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics **11**(1), 395 (2010)

4. Tautenhahn, R., Bottcher, C., Neumann, S.: Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics **9**(1), 504 (2008). doi:10.1186/1471-2105-9-504

5. Conley, C.J., Smith, R., Torgrip, R.J., Taylor, R.M., Tautenhahn, R., Prince, J.T.: Massifquant: open-source Kalman filter based XC-MS isotope trace feature detection. Bioinformatics **30**(18), 359 (2014)

6. Smith, R., Ventura, D., Prince, J.T.: Novel algorithms and the benefits of comparative validation. Bioinformatics **29**(12), 1583–1585 (2013)

7. Rosen, J., Handy, K., Gillan, A., Smith, R.: Js-ms: a cross-platform, modular javascript viewer for mass spectrometry signals. BMC Bioinformatics **18**(1), 469 (2017)

8. Handy, K., Rosen, J., Gillan, A., Smith, R.: Fast, axis-agnostic, dynamically summarized storage and retrieval for mass spectrometry data. PLOS One **12**(11) (2017)

9. Henning, J., Tostengard, A., Smith, R.: A peptide-level fully annotated dataset for quantitative evaluation of precursor-aware mass spectrometry data processing algorithms. Journal of Proteome Research **18**(1) (2018)

10. Gutierrez, M., Handy, K., Smith, R.: Quantitative evaluation of algorithms for isotopic envelope extraction via extracted ion chromatogram clustering. Journal of Proteome Research **17**(11) (2018)

11. Danchilla, B.: Three.js framework. In: Beginning WebGL for HTML5, pp. 173–203. Springer, N.p. (2012)

12. Smith, R., Ventura, D., Prince, J.T.: Controlling for Confounding Variables in MS-omics Protocol: Why Modularity Matters. Briefings in Bioinformatics **15**(5), 768–770 (2014)

13. Nahnsen, S., Bertsch, A., Rahnenf ¨uhrer, J., Nordheim, A., Kohlbacher, O.: Probabilistic consensus scoring improves tandem mass spectrometry peptide identification. Journal of Proteome Research **10**(8), 3332–3343 (2011)

A Peptide-Level Fully Annotated Dataset for Quantitative Evaluation of Precursor-Aware Mass Spectrometry Data Processing Algorithms

Jessica Henning, Annika Tostengard, and Rob Smith

Abstract

Modern label-free quantitative mass spectrometry workflows are complex experimental chains for devising the composition of biological samples. With benchtop and in silico experimental steps that each have a significant effect on the accuracy, coverage, and statistical significance of the study results, it is crucial to understand the efficacy and biases of each protocol decision. While many studies have been conducted on wet lab experimental protocols, post-acquisition data processing methods have not been adequately evaluated, in large part due to a lack of available ground truth data. In this manuscript, we provide a novel ground truth dataset for mass spectrometry data analysis at the precursor (MS1) signal level comprised of isolated peptide signals from UPS2, a popular complex standard for proteomics analysis, requiring more than 1,000 hours of manual curation. The dataset consists of more than 62 million points, with 1,294,008 grouped into 57,518 extracted ion chromatograms, and those grouped into 14,111 isotopic envelopes. This dataset can be used to evaluate many aspects of mass spectrometry data processing, including precursor mapping and signal extraction algorithms.

Introduction

Mass spectrometry technology is vital for answering a variety of experimental questions across many disciplines [1]. Mass spectrometry plays a role in many biological and biomedical investigations because it can quantify and identify the major components (proteins, lipids, metabolites) of almost any cellular system. Mass spectrometry creates raw output that, when analyzed with data processing software, can be used to elucidate the identities and quantities of molecules in the analyzed sample.

Existing workflows have been used to produce an astonishing volume of publications and discoveries. Still, quantitative evaluations of data processing approaches are sparse [2], and practitioners regularly encounter data processing results that raise questions regarding the accuracy of current workflows.

Not knowing which methods perform well or poorly and under what conditions limits the accuracy and impact of scientific discovery [3]. Practitioners require a knowledge of the performance of the algorithms they consider for data processing, as the choice of algorithm can affect the experimental outcome as much as wet lab protocol choice [4].

Moreover, scientific studies require reporting of uncertainty and accuracy measures in order to inform the broader audience of the limitations of the study results. Yet, the uncertainty inherent in data processing steps is not necessarily reflected by the current metrics (e.g. target-decoy false detection rate estimation) widely employed in mass spectrometry today [5, 6]. Quantitative evaluations will provide insight on the real-world performance of common and novel approaches, and can provide observations to improve Bayesian-based estimates of uncertainty recently proposed [7].

Across science, benchmark datasets are used to quantitatively evaluate existing approaches–to assess strengths and weaknesses and inform the development of next generation solutions. In mass spectrometry, many benchmark datasets have been published (for example [8, 9, 10]). However, the need for more and better ground truth is still acute.

LC-MS experiments result in raw output, typically consisting of precursor (MS1) and fragmentation (MS2) data, that must be analyzed with data processing software to yield molecular identities and quantities. In spite of the fact that the choice and application of data processing software can have as dramatic an effect on experimental results as benchtop protocol, very few algorithms and software have been quantitatively evaluated.

Benchmark datasets are used to quantitatively evaluate existing approaches to assess strengths and weaknesses and inform the development of the next generation solutions. In mass spectrometry, many benchmark datasets have been published, but there is still a need for more and better ground truth.

First, while existing “ground truth” datasets are a necessary bootstrap to the ideal, they are limited in their capabilities. Typically, the literature uses the phrase ”ground truth” to describe experimental datasets that contain spiked-in standards [10]. Meanwhile, the word ”comprehensive” is usually used to indicate that the dataset has been analyzed as many times as possible with current analytical methods. Current ground truth methods are not an external, objective measurement but at best the intersection of overlapping sets of the differing results from the very methods under evaluation [11].

Second, these datasets are end-to-end (for example [12, 8]), meaning that they are intended to test the full data processing pipeline that renders raw data into molecular identities and quantities. There are many steps in the mass spectrometry data processing pipeline. Some are a critical piece of standard workflows, such as the interpretation of tandem mass spectrometry spectra for molecular identification. Others are optional. Specific attribution of error is important given that each module in the data processing pipeline can have a significant effect on the overall result [4, 13]. While end-to-end ground truth datasets are valuable for demonstrating performance of entire workflows, they are incapable of evaluating the influence of the choice of algorithm and parameter settings for each module, precluding identification and resolution of performance bottlenecks. Developers of new algorithms and practitioners alike require modular benchmark datasets to guide algorithm choice and algorithmic development–datasets that focus on specific steps in the data processing pipeline. Recently, the community has increasingly called for modular benchmark datasets [14, 2, 11, 15, 13].

While in silico simulation provides the molecular provenance of every point in the mass spectrometry output, the complexity and current observational limitations of mass spectrometry data has caused many in the community to have major reservations about using simulated data for algorithmic evaluations [13]. Though several programs have been published for simulating mass spectrometry data [16, 17, 18, 19], enabling the inexpensive generation of any number of ground truth datasets covering a broad set of experimental scenarios, they have not been adopted by the community and probably will not be until sufficient real world ground truth is produced to validate and improve them.

Perhaps due to these limitations in current ground truth data, novel mass spectrometry data processing algorithms are often not subjected to the thorough comparative evaluation endured by their counterparts in other science fields, where an algorithm unevaluated against extant methods is considered unpublishable. For example, Smith et al. showed that for the specific problem of mass spectrometry data alignment, more than half of the novel algorithms published failed to show an evaluative comparison to even one of the more than 100 published alternatives [2].

An ideal benchmark would be representative–consisting of sufficient individual datasets to provide a high fidelity sample of the subtle and diverse characteristics of real data. Boulesteix et al. make a cogent argument that a representative comparison study is an extremely difficult and time-intensive task [14]. We present the following work as a first step towards this end. In its present form, the dataset provided here can be used to provide illustrative pilot evaluations on mass spectrometry data processing modules that rely on MS1 input to quantify the capabilities of existing or new algorithms on a real dataset.

One common approach to evaluating mass spectrometry data processing algorithm accuracy without the benefit of a ground truth dataset is to use consensus results. For example, if one desires to know how well an XIC-extraction method performs, they might compare the result list to that of another XIC-extraction algorithm. These consensus evaluations are qualitative, not quantitative. While qualitative evaluations are common, these are inherently limited to answering the question, “how well do new methods perform compared to old methods?” This is not a particularly valuable question to ask, given that, without being quantitatively evaluated, the original method’s results are no more likely to be correct than the new method’s. In reality, the question of interest is “how well do new methods perform compared to the true answer?” This latter question can only be answered using ground truth: data for which the true answer is known.

Quantitative evaluations provide insight on the real-world performance of common approaches, highlight persisting weaknesses, and provide direction for novel approaches. Quantitative evaluations also provide a quality control barrier of entry for novel methods, helping to mitigate the proliferation of publications that make it difficult for practitioners to keep track of the state of the art. For example, a recent review of LC-MS alignment methods showed that very few of more than 50 published algorithms provided novel functionality compared to previously published methods.

In this manuscript, we describe a new, fully-annotated quantitative ground truth MS1 dataset. Using a new open-source software, JS-MS 2.0, we conducted an extensive manual annotation process comprised of more than 1,000 human hours. The result is an MS1 ground truth dataset designed to evaluate mass spectrometry data processing algorithms that operate on MS1 data. The dataset consists of more than 62 million points, with 1,294,008 grouped into 57,518 extracted ion chromatograms, and those grouped into 14,111 isotopic envelopes. This dataset can be used to evaluate many aspects of mass spectrometry data processing, including precursor mapping and signal extraction algorithms. To demonstrate that quantitative ground truth provides more utility than consensus results-based evaluation, we show an example evaluation of several popular XIC-extraction algorithms on one window of data from the dataset.

Methods

The curated dataset is an untargeted protein identification sample consisting of 48 Universal Proteomics Standard 2 (UPS2) proteins. UPS2 has been used in many publications as a known set of molecules and abundances that approaches the large dynamic range of abundances present in naturally-occurring biological samples [20, 21, 22]. The proteins are organized into six groups of abundances with eight protein types per group. The abundances per group vary from 0.5 fmol to 50,000 fmol with each group differing by an order of magnitude.

Raw Data

The raw data consists of a trypsin-digested run of UPS2 produced and recently published by the Nesvizhskii group as part of comparison of state-of-the-art data-dependent and data-independent acquisition methods [12]. It provides an independent representative example of a run using modern instrumentation, wet lab, and instrumental protocol. The file in question was created using a data independent acquisition protocol on an AB Sciex TripleTOF 5600, using a 250-ms ion accumulation time for MS1 survey scans. The raw data file is publicly available in PRIDE repository PXD001587 under filename 18185_REP2_4pmol_UPS2_IDA_1.mzXML and consists of data centroided by ProteinPilot (Sciex) software. The mzXML file was converted to mzML using msconvert [23].

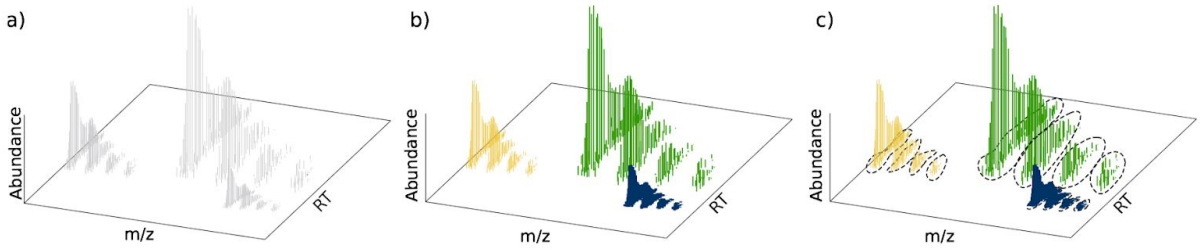


Figure 1: MS1 Data Segmentation. Proteomics mass spectrometry raw precursor data consists of many points generated by (a) the detection of many peptides of a given composition and charge state. One way to segment raw points into isotopic envelopes is by (c) first grouping them into the extracted ion chromatograms (XICs) caused by isotopic variants, then clustering them into envelopes (b).

Software

The ground truth methods described below are not uniquely associated with any particular piece of software. While software facilitated in the visualization of 2-d and 3-d views of the data, and provided easy interfaces for displaying, modifying, and capturing user data segmentations, no computational decision processes (algorithms) were used, only manual decisions by the users. The software used was JS-MS 2.0, a new open-source software for mass spectrometry data viewing and annotation, by the Smith Lab. Like other tools such as OpenMS TOPPView, JS-MS 2.0 allows users to view 2-d and 3-d representations of data. Unlike other tools, it has advanced navigation abilities and interfaces for quickly delineating, storing, and editing MS1 signal boundaries in mass spectrometry data.

Ground Truth Curation Process

Mass spectrometry data is generated through the detection of ionized molecules which register intensities at their mass-to-charge ratio (m/z) and retention time (RT) (see Figure 1a). Each type of ionized molecule in a sample will create a unique 3-d signal group, called an isotopic envelope, at a specific m/z and RT for every molecule at every charge state (see Figure 1b). Each envelope is comprised of one or more extracted ion chromatograms manifesting the molecular differences caused by natural (e.g. [11] C vs. [12] C) or artificial (e.g. deuterated water) isotopic variations (see Figure 1c).

First, all signals were segmented into extracted ion chromatograms (XICs) as follows: For each region, XICs were delineated by rise out of background and fall back into background where interleaving points demonstrated low m/z variance, with greater variance allowed for lower intensity signals, and a roughly Gaussian shape (with lower apex expected for lower abundance signals). Peak intensity variation within XICs was allowed to the extent that it did not violate an on-average Gaussian shape as judged by the operator.

Putative XICs were split in the RT direction when overlapping chromatographic distributions were revealed using the 2-d and 3-d views provided by JS-MS 2.0 (see Figure 2b). In the given dataset, some poor centroiding results created unusually skewed raw data (see Figure 2c), and 0.001 Da was used as a cutoff for m/z deviation in these cases.

After all points were either classified as noise or clustered into XICs (see Figure 2a), XICs in each region were clustered into isotopic envelopes. XICs that co-eluted, displayed similar intensity distributions, and occurred at fixed m/z distances from each other were grouped into isotopic envelopes. Valid m/z distances were restricted to 1/n, where n is a real positive integer. In cases where XICs overlapped, they were deconvolved through comparing intensity apex in RT, m/z distances, and similarity of intensity distributions.

Points with low enough intensity to preclude distinction from surrounding background were marked as noise (see Figure 3b).

Deconvolution

In the event of multiple XICs overlapping in m/z and RT, the lower abundance ions were used to delineate a splitting point for the higher abundance, overlapping XICs. A hard, point-resolved boundary was enforced for each signal. It should be noted that this is one weakness in the approach. While enforcing points' membership in only one trace prevents the more accurate splitting of point intensities between multiple traces when those traces overlap, it allows the dataset to be independent of biases that would come from applying any one deconvolution algorithm. In practice, many signals do not overlap to the degree that underlying signals would be considerably affected.

Validation

After initial segmentation, the m/z, and RT coordinates of each envelope were combined to create a worklist for validation. Each isotopic envelope target and the region around it were re-segmented by a single senior technician. In the event of disagreement, the senior technician's segmentation was retained. In this way, each envelope in the segmentation was analyzed twice–once in the original segmentation, and once by an independent senior technician.

Output Format

The curation process assigns every point in the file as either a member of an XIC or noise. Each XIC, in turn, is assigned to an isotopic envelope.

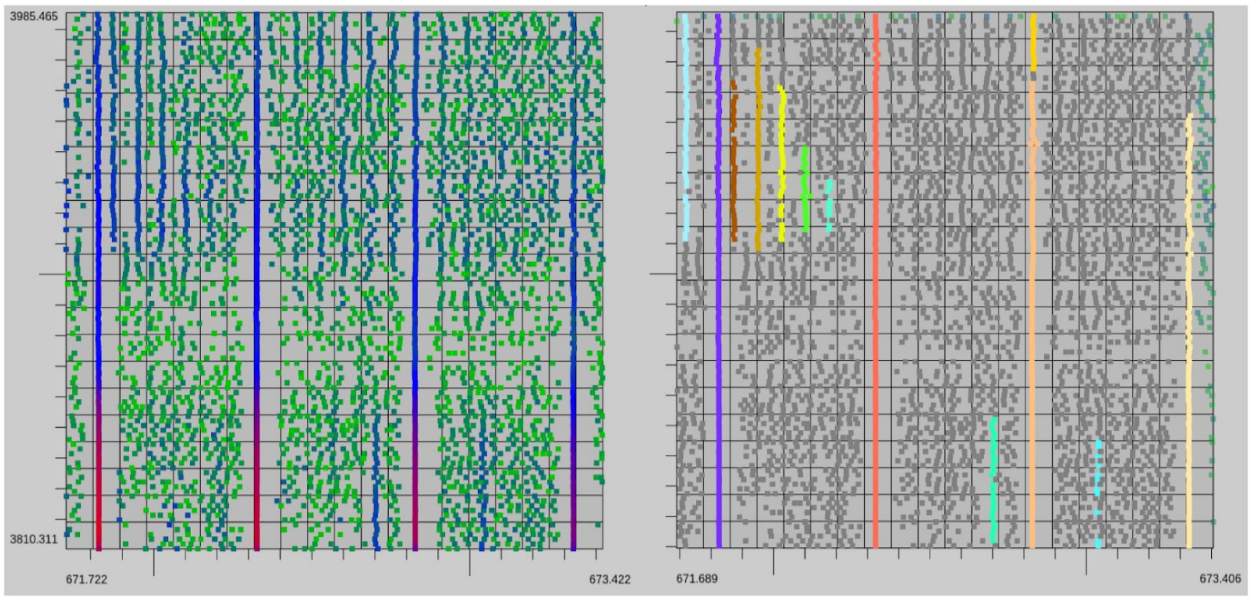
In order to be useful to the greatest number of investigators, the data is freely available in a .csv format with one row per data point, and the following schema:

point id, m/z, RT, intensity, XIC id, isotopic envelope id

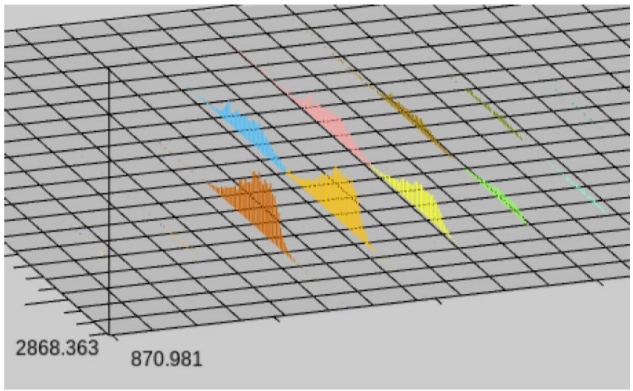
where point id is a unique identifier and XIC id is set to -1 for noise points.

Example Evaluation

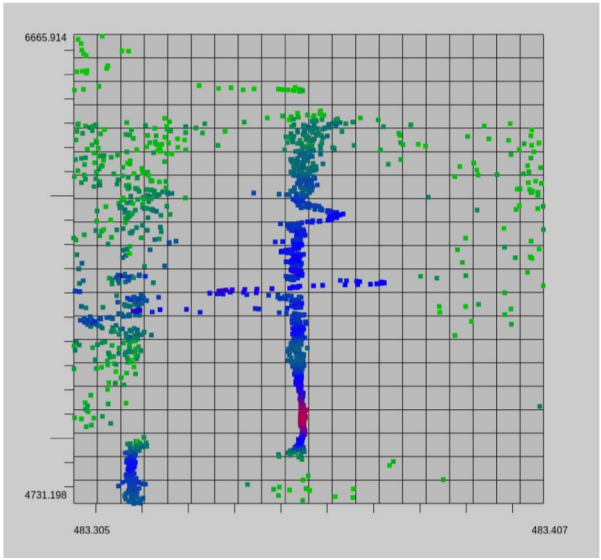
In order to illustrate the additional information provided by quantitative ground truth over qualitative evaluation (e.g., consensus results), we conducted a brief XIC evaluation on several popular algorithms for the task from the XCMS library, including CentWave [24], MatchedFilter [24], and Massifquant [25]. Although a full evaluation of these methods on the entire dataset is beyond the scope of this paper, this analysis is sufficient to motivate the need for quantitative ground truth. In this analysis, each algorithm's default parameters were used as a baseline. Additionally, the Isotopologue Parameter Optimization tool [26] (IPO) was used to find optimal results for CentWave and MatchedFilter. IPO does not provide an optimization process for Massifquant, and an exhaustive parameter search was beyond the scope of this paper.



(a) Noise Assignment. Raw data in 2-d (left, colors indicate intensity) and segmented data (right, colors indicate XIC membership). After all distinguishable signals were segmented, the remaining points were marked as noise (gray points, right panel).



(b) Chromatographic Overlaps. Putative XICs were split point wise through evaluation of visualized overlap in 2- and 3-d views.



(c) Poor Centroiding. In locations with poor centroiding, 0.001 Da was used as a m/z deviation limit.

Figure 2: Data was segmented using JS-MS 2.0

Results

The original UPS2 file has over 62 million points. Over 1.2 million of these presented with sufficient information to be segmented, with a significant skew towards lower intensity points (see Figure 3a, note the logarithmic scale of the y axis). Not surprisingly, the vast majority of points that did not present sufficient information to be segmented were low intensity, but a surprising number of points of high intensity did not provide sufficient information to be segmented into XICs (see Figure 3b).

Over 57,000 XICs were segmented (see Figure 3c, note the linear scale). While one might predict that most segmentable XICs would be high intensity, the data suggests otherwise. The XICs were clustered into 14,111 envelopes, each presumably corresponding to a unique peptide/charge state combination, or several in the case of isomers (see Figure 3d, note the logarithmic scale). As one might expect, there are far more low intensity envelopes than high intensity envelopes.

Discussion

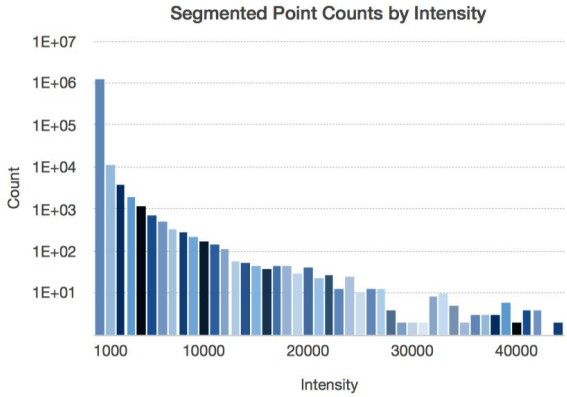
This quantitative ground truth dataset is designed to advance the capability of quantitative evaluations in mass spectrometry data processing.

The principle purpose of this dataset is to serve as a basis for quantitative evaluations of algorithms that involve or rely upon MS1 data. This dataset can be used to quantify evaluations of:

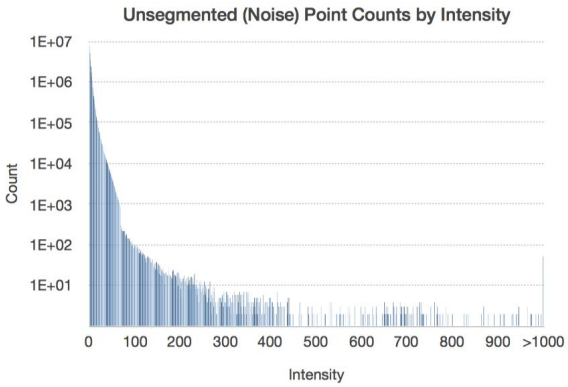
- XIC extraction algorithms.
- Algorithms that cluster XICs into isotopic envelopes.
- MS1-based quantification methods.
- MS2-based quantification methods.
- False detection estimations.

Potential Application: Quantitative Validation

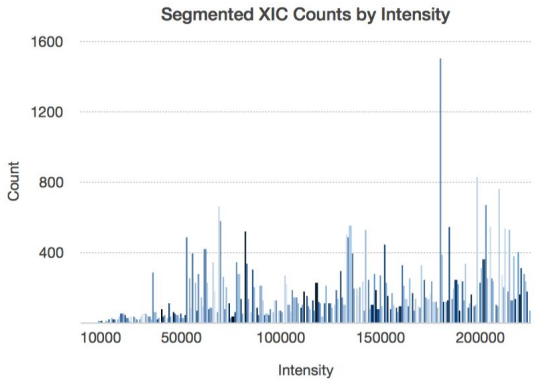
There are several uses for ground truth data. One is illustrated in Figure 4. Here, one small window of the run (see Figure 4a) is shown, as well as the same tile after manual ground truth segmentation (see Figure 4b). In the other tiles, segmentations provided by several algorithms within XCMS are provided. By inspecting each subfigure, the viewer will note differences in the point membership of each XIC between algorithms. If a user seeking to evaluate the performance of a set of algorithms for a small window of data such as this, they could imagine what the correct segmentation looks like and, in this case, choose Massifquant or CentWave for their data (probably depending on time available, as CentWave is much faster to run). But without quantitative metrics, how can you make the same decision across the scope of a whole run? Unfortunately, without a validated ground truth, quantitative comparisons are not possible. With ground truth available, metrics like m/z accuracy and intensity accuracy can be reported, making it possible to measure algorithm performance with respect to reality as opposed to with respect to other algorithms.



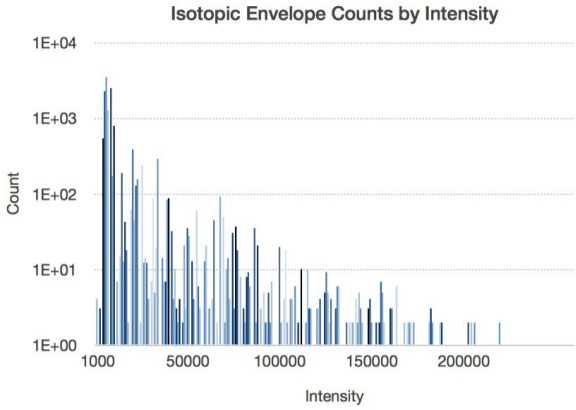
(a) The ground truth dataset contains segmentations for 1,294,008 of the over 62 million points in this file.



(b) The vast majority of points in this dataset were very low intensity. Most did not provide sufficient information to yield a segmentation, and were therefore classified as noise.



(c) The ground truth contains 57,518 XICs, with a subtle skew towards higher intensity signals.



(d) The ground truth data contains 14,111 unique isotopic envelopes. A super majority of envelopes have very low intensity

Figure 3: Dataset properties

In addition to quantitative evaluations, it is hoped that this and future contributions will highlight additional applications for and impacts of using MS1 data.

Study Weaknesses

In practice, XICs of multiple envelopes can occur at unresolved m/z, as can be seen in Figure 4a where many more XICs exist than reflected in the ground truth segmentation (see Figure 4b, as at least four envelopes overlap in the center cluster of points). In the protocol employed in this study, points were only allowed to pertain to a single XIC. While this was an intentional choice to avoid subjective decisions when splitting the intensity of shared points between multiple XICs, a more accurate ground truth estimate could be provided if a perfect intensity splitting algorithm were available. It is important to note that none of the XIC extraction algorithms applied to the tile in Figure 4 were able to handle overlapping XICs. While this dataset is not perfectly segmented, it is important to note that 1) it provides a significant improvement over qualitative ground truth, 2) it provides an intermediate bootstrap step to something better, 3) these limitations matter more for isotopic envelope segmentation evaluation than XIC segmentation.

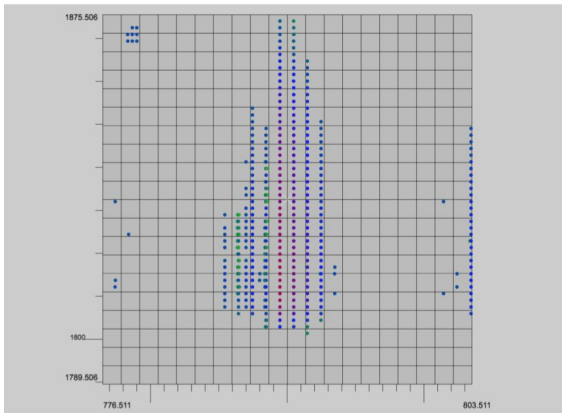
Another weakness of this dataset is the fact that the raw data is centroided. Current centroiding algorithms are far from perfect, and incorrect centroiding destroys data that could be used to render a more accurate XIC or envelope segmentation. However, since many users employ centroided data, this dataset reflects the reality of many workflows.

Potential Application: MS1-based Identification

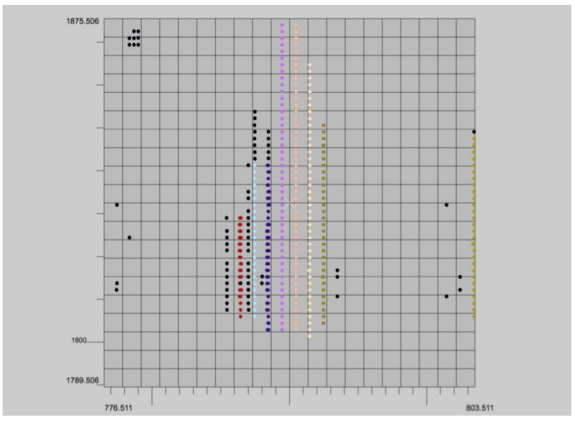
One such application may be an increased number of peptide identifications using protocols that combine MS1 and MS/MS. The original study providing this dataset demonstrated that state of the art DDA and DIA methods were able to identify 9,272 and 8,757 peptides (respectively) from this dataset using typical MS/MS-only identification algorithms [12]. This dataset contains more than 14,000 annotated isotopic envelopes from a single run. Many of these envelopes are low abundance (see Figure 3d), suggesting that MS1-based approaches for identification may yield improved coverage and lower limit of detection compared to MS/MS-based approaches alone. One possible way of combining the information provided by each approach is by designing a protocol that uses a single MS1 run and off-instrument feature detection to enumerate a target list for successive DDA replicates. In this way, all 14,000 envelopes could be targeted in a principled way. MS/MS identification could be performed on the spectra obtained, while the MS1 envelopes could provide additional information via charge state, isotope relative ratio, and precursor m/z for validation and/or to increase the specificity of the MS/MS reference database.

Future Work

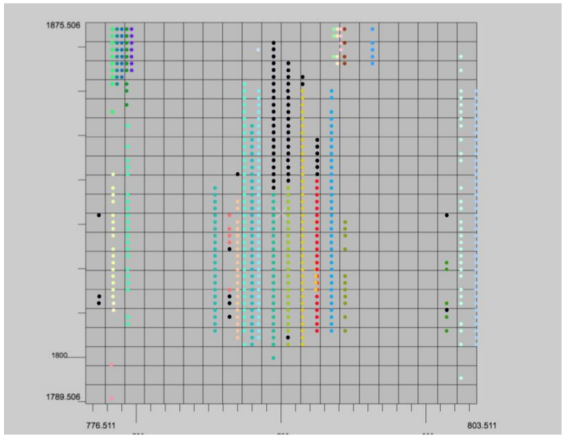
Future work will consist of creating more datasets, including multiple replicates that can be used to evaluate tasks such as retention time alignment. Our group is also conducting quantitative evaluations on downstream algorithms, such as XIC extraction, isotopic envelope finding, and MS1-MS/MS signal mapping.



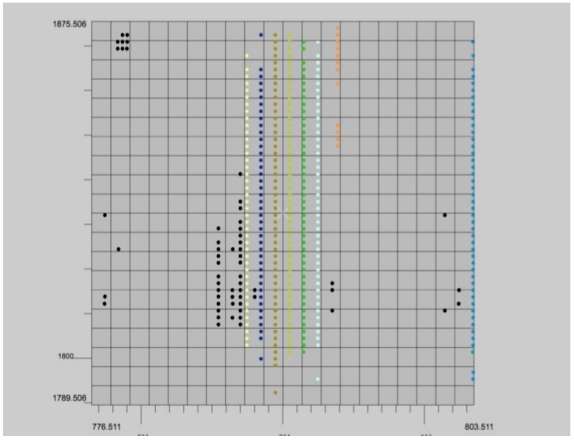
(a) One tile of unsegmented raw data.



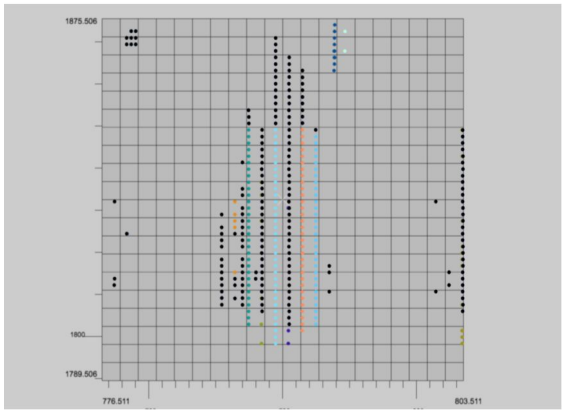
(b) Manual XIC segmentation (ground truth).



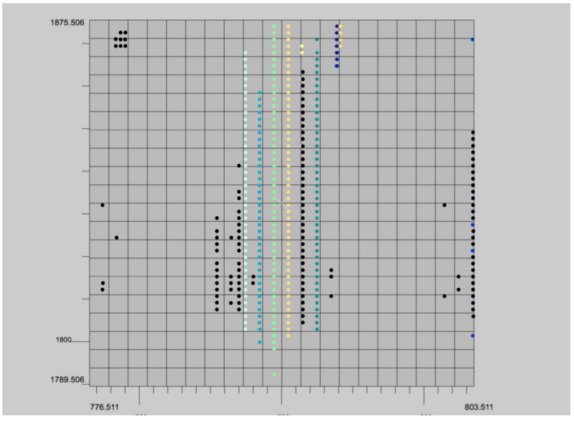
(c) XCMS's CentWave (default settings).



(d) XCMS's CentWave (optimized settings).



(e) XCMS's matchedFilter (optimized settings).



(f) XCMS's Massifquant (default settings).

Figure 4: Ground truth allows the quantitative comparison of output. Here, the color of each point shows XIC membership (black points were not assigned) as determined by various algorithms in panels other than a), where the color of each point corresponds to its intensity.

Conclusion

Many mass spectrometry data processing algorithms have been proposed that rely on MS1 data, including XIC extraction algorithms, algorithms that cluster XICs into isotopic envelopes, and MS1-based quantification methods. In spite of the number of these methods proposed, none have been evaluated based on quantitative ground truth. Qualitative comparisons based on consensus results have not adequately demonstrated algorithm performance.

In this manuscript, we describe a new quantitative dataset for MS1 data. With more than 1,000 hours of manual curation, the dataset contains annotations of more than 62 million points, with 1,294,008 grouped into 57,518 extracted ion chromatograms, and those grouped into 14,111 isotopic envelopes. This dataset can be used to evaluate many aspects of mass spectrometry data processing, including precursor mapping and signal extraction algorithms.

Availability and License: The dataset is available on Github (https://github.com/optimusmoose/ups2GT) with a non-commercial license. For a commercial license, please contact the author.

Author Contributions
J.H. led the data curation effort. A.T. conducted the evaluation experiments. R.S. designed and supervised the project. All authors wrote the manuscript.

Funding
This research was supported by NSF grant number 366208 to R.S.

Materials & Correspondence
should be directed to R.S.

References

[1] Griffiths, W. J. & Wang, Y. Mass spectrometry: from proteomics to metabolomics and lipidomics. *Chem. Soc. Rev.* **38**, 1882–1896 (2009). URL http://dx.doi.org/10.1039/B618553N.

[2] Smith, R., Ventura, D. & Prince, J. T. Novel algorithms and the benefits of comparative validation. *Bioinformatics* **29**, 1583–1585 (2013).

[3] Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data the protein inference problem. *Molecular & Cellular Proteomics* **4**, 1419–1440 (2005).

[4] Smith, R., Ventura, D. & Prince, J. T. Controlling for Confounding Variables in MS-omics Protocol: Why Modularity Matters. *Briefings in Bioinformatics* (2013).

[5] Serang, O. & Noble, W. A review of statistical methods for protein identification using tandem mass spectrometry. *Statistics and its interface* **5**, 3 (2012).

[6] Serang, O. & Kall, L. Solution to statistical challenges in proteomics is more statistics, not less. *Journal of proteome research* **14**, 4099–4103 (2015).

[7] Serang, O., Moruz, L., Hoopmann, M. R. & Kall, L. Recognizing uncertainty increases robustness and reproducibility of mass spectrometry-based protein inferences. *Journal of proteome research* **11**, 5586–5591 (2012).

[8] Ramus, C. *et al*. Benchmarking quantitative label-free lc–ms data processing workflows using a complex spiked proteomic standard dataset. *Journal of Proteomics* **132**, 51–62 (2016).

[9] Franceschi, P., Masuero, D., Vrhovsek, U., Mattivi, F. & Wehrens, R. A benchmark spike-in data set for biomarker identification in metabolomics. *Journal of chemometrics* **26**, 16–24 (2012).

[10] Wessels, H. J. *et al*. A comprehensive full factorial lc-ms/ms proteomics benchmark data set. *Proteomics* **12**, 2276–2281 (2012).

[11] Allmer, J. A call for benchmark data in mass spectrometry-based proteomics. *Journal of Integrated OMICS* **2**, 1–5 (2012).

[12] Tsou, C.-C. *et al*. Dia-umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nature methods* **12**, 258–264 (2015).

[13] Gatto, L. *et al*. Testing and validation of computational methods for mass spectrometry. *Journal of proteome research* **15**, 809–814 (2015).

[14] Boulesteix, A.-L. On representative and illustrative comparisons with real data in bioinformatics: response to the letter to the editor by smith et al. *Bioinformatics* btt458 (2013).

[15] Noble, W. S. & MacCoss, M. J. Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput Biol* **8**, e1002296 (2012).

[16] Bielow, C., Aiche, S., Andreotti, S. & Reinert, K. MSSimulator: Simulation of mass spectrometry data. *Journal of Proteome Research* **10**, 2922–2929 (2011).

[17] Smith, R. & Prince, J. T. Jamss: proteomics mass spectrometry simulation in java. *Bioinformatics* **31**, 791–793 (2015).

[18] Noyce, A. B. *et al*. Mspire-Simulator: LC-MS Shotgun Proteomic Simulator for Creating Realistic Gold Standard Data. *Journal of Proteome Research* **12**, 5742–5749 (2013).

[19] Prince, J. T. & Smith, R. Probabilistic generation of mass spectrometry molecular abundance variance for case and control replicates. *Journal of proteome research* **16**, 2429–2434 (2017).

[20] Soufi, B., Krug, K., Harst, A. & Macek, B. Characterization of the e. coli proteome and its modifications during growth and ethanol stress. *Frontiers in microbiology* **6**, 103 (2015).

[21] Bogdanovi´c, O. et al. Active dna demethylation at enhancers during the vertebrate phylotypic period. *Nature* **201**, 6 (2016).

[22] Hubner, N. C., Nguyen, L. N., Hornig, N. C. & Stunnenberg, H. G. A quantitative proteomics tool to identify dna–protein interactions in primary cells or blood. *Journal of proteome research* **14**, 1315–1329 (2015).

[23] Adusumilli, R. & Mallick, P. Data conversion with proteowizard msconvert. In *Proteomics*, 339–368 (Springer, 2017).

[24] Tautenhahn, R., Bottcher, C. & Neumann, S. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* 9, 504 (2008). URL http://www.biomedcentral.com/1471-2105/9/504.

[25] Conley, C. J. et al. Massifquant: open-source Kalman filter-based XC-MS isotope trace feature detection. *Bioinformatics* **30**, 2636–2643 (2014).

[26] Libiseller, G. *et al*. Ipo: a tool for automated optimization of xcms parameters. *BMC Bioinformatics* **16**, 118 (2015).