# *mineXpert*: biological mass spectrometry data visualization and mining with full JavaScript-ability

Filippo Rusconi[*,†,‡]

†*LCP, UMR CNRS 8000, Université Paris-Sud, Université Paris Saclay. Build. 349, 91400 Orsay, France*

‡*PAPPSO, GQE–Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190, Gif-sur-Yvette, France*

E-mail: filippo.rusconi@u-psud.fr

Phone: +33 (0)1 69 33 23 54. Fax: +33 (0)1 69 33 23 40

**Abstract**

Biological mass spectrometry mainly comprises three fields of endeavour, namely, proteomics, metabolomics and structural biology. In each of these specialties, the mass spectrometrist needs to access $MS^1$ mass spectral data, although not necessarily on the same basis. For example, the bottom-up proteomics scientist will occasionally access $MS^1$ data in order to perform data inspection, quality assessments and quantitation measurements, while top-down proteomics, structural biology or metabolomics scientists will actually spend most of their time mining profile-mode $MS^1$ data. Furthermore, the advent of ion mobility–mass spectrometry imposes new manners of mass spectral data visualization. An open-source $MS^1$-only mass data visualization software for the desktop was developed to allow scientists to visualize conventional and drift time mass data. Various mass data integrations are possible, allowing a thorough mass spectral

data scrutiny. Isotopic cluster calculations are easily carried over from the chemical formula up to the display of the mass spectrum. Deconvolution of mass peaks can be achieved with a simple mouse drag. Flexible reporting of data inspection events and of mining discoveries is provided. Very large sparse datasets can be sliced into smaller chunks replicating the original data without data loss. Task automation is achieved in a JavaScript environment. This project allows users of mass spectrometry facilities to inspect and mine their $MS^1$ mass data outside of these facilities, without having to resort to the closed-source vendor software shipped with the instruments. *mineXpert* requires no proprietary software whatsoever once the mass spectrometry data have been converted to mzML. The reference implementation is version 5.8.2 or greater. Reference material, a detailed user manual and video tutorials are available at `http://www.msxpertsuite.org`.

## Keywords

mass spectrometry, ion mobility, MS1 spectrum, data visualization, proteomics, isotopic cluster, metabolomics, software, C++, JavaScript, mzML, GPLv3+.

## Introduction

Mass spectrometry is an essential technique in the field of biology as it allows scientists to probe biological systems at different levels: the structure of isolated analytes (*e.g.* structural biology); the identification and quantitation of analytes in complex mixtures (*e.g* proteomics and metabolomics). In bottom-up proteomics and, to a lesser extent, in metabolomics, the sheer number of analytes in any given sample requires software to almost automatically process the data (OpenMS,[1] ProteoWizard,[2] X!TandemPipeline,[3] massChroQ[4]). In top-down proteomics or structural biology mass spectrometry, the mass data are acquired on enriched or purified analytes with the aim of identifying them and characterizing their fine

structure (*e.g.* chemical modification, covalent adduct, cross-link, non covalent polymer-ligand interactions). Whatever the field of endeavour, the biological mass spectrometrist will invariably access the $MS^1$ mass data (*i.e.* survey scans), either to perform quality assessments, quantitation measurements or to actually mine the data. In the specific case of bottom-up proteomics, in the eventuality of dubious protein identifications, resorting to the visualization of precursor ions in the original survey scans is common practice, as the shape or intensity of m/z peaks can enlighten the scientist on the validity of a peptide-spectrum match. These tasks are highly time-consuming, making an efficient mass data visualization and mining tool highly desirable, even more so nowadays, as mass spectrometry data have ever greater depths: hyphenated mass spectrometry techniques produce massive amounts of time-resolved mass spectral datasets and the recent advent of ion mobility–mass spectrometry (IM–MS) has added another extra layer of time-resolved spectral data. Finally, the technical validation of mass spectral data acquisitions is invariably performed by looking manually at the acquired data. In medium-sized datasets that are typical of the average mass spectrometry facility, this task becomes tedious and error prone. Therefore, the ability to automate the various steps involved in the technical validation, like data file loading, mass spectral integration, statistical calculations over the total ion current (referred to as TIC in this report), becomes an appealing feature.

Several biological mass spectrometry data visualization software projects exist. MZmine 2 and BatMass provide a Java-based software environment but do not support IM–MS data.[5,6] The mMass software, a Python-based proteomics-oriented spectrum viewer, has no IM–MS capabilities[7]. Amphitrite provides a Python-based IM–MS-centric software solution that lacks a flexible mass spectrum viewer capable of loading an arbitrary number of datasets simultaneously[8]. TOPPView[1] and SeeMS[2] both allow highly detailed mass spectral data scrutiny but have no combination/integration features. Skyline, also a free software program, is mainly aimed at simplifying the elaboration of mass data acquisition methods and the analysis of data from targeted proteomics experiments[9]. To the best of my knowledge, none

of the software above can be automated using a scripting language.

*mineXpert* is a JavaScript-able desktop software program implementing fluid navigation through the full depth of TIC and extracted ion current (XIC) chromatograms, $MS^1$ and drift time mass spectrometric data. Users of mass spectrometry facilities have to adapt to the various mass spectrometry software programs that are shipped with the instruments, making their data visualization and mining an unnecessarily difficult task. *mineXpert* implements all the commonest features found in a variety of proprietary software programs so that a single program will fit all the needs, whatever the instrument whence the data originated. The scripting capabilities of *mineXpert* allow scientists to automate and reproduce data visualization and mining procedures. Data mining procedures include easy deconvolution of mass peaks, recording of useful features and reporting of mining discoveries in expressive ways. *mineXpert* is bundled with the *massXpert*[10,11] polymer modelling and mass spectrometry simulation software to form the *msXpertSuite* software project (`www.msxpertsuite.org`). The whole software suite, datasets and documentation files are licensed under the GNU General Public License version (GPL v3+).

# Methods

The software is written in C++11 using the cross-platform Qt libraries (`www.qt.io`). The software is thus compatible with MS Windows (tested versions: 7 and 10), Linux and MacOSX (versions 10.11.6 and newer). The QCustomPlot library (`www.qcustomplot.com`) is used for plotting. Code parallelisation is based on OpenMP (`www.open-mpi.org`), a *de facto* standard that implements portable parallel code at compile time under the control of compiler directives in the code. Tests revealed that parallelized code shows execution speed gains almost proportional to the number of processing threads.

The mass spectrometry data file loading process is handled by the libpwiz library (ProteoWizard) and the isotopic cluster calculations are performed using the libIsoSpec++ library.[12] Both

the libpwiz and libIsoSpec++ libraries are expressly built by the author for all the supported platforms and included in all the binary packages. The main development platform is Debian GNU/Linux (`www.debian.org`). As an official Debian Developer, the author produces software packages for Debian/Ubuntu and derivatives (typically installed using the `apt-get install msxpertsuite` command). Binary packages are available for MS Windows and MacOSX (`www.msxpertsuite.org`). *mineXpert* has been tested with ion mobiliy mass spectrometric data files from the Synapt G2Si-HDMS instrument and from the Agilent 6560 ESI-IMS-Q-TOF instrument.[13] Conventional mass spectrometry data files could be loaded from a wide variety of mass spectrometers.

*mineXpert* loads mass data files in the mzML format from the Proteomics Standards Initiative[14] (see[15] for a thorough description), as obtained by conversion of raw data files using the msconvert[16] tool in the ProteoWizard software package. When mass spectrum scrutiny is required for a unitary mass spectrum, *mineXpert* loads any data in the generic <m/z><separator><intensity> text format with <separator> being any non-numeric character sequence. A single mass spectrum can be loaded from file or directly from the clipboard. The mzML format was selected because it is a standard supported on all the computing platforms. However, in future versions and on the MS Windows platform, the software might be able to load data directly from proprietary raw files using the libpwiz library.

For this report, the dataset that is used to describe *mineXpert* was from a size-exclusion chromatography–IM–MS study of the bovine alcohol dehydrogenase in native conditions; 6 Gb dataset containing 318462 non-empty spectra.[17]

# Results and discussion

The capabilities of the *mineXpert* software will now be described in detail following a functional path from data file loading to the recording of mining discoveries.

## Graphical user interface fundamentals

*mineXpert* allows one to load any number of mass data files in the same session. The general approach of the *mineXpert* software architecture is to allow mass spectral scrutiny in dedicated windows (Figure S1) that get progressively populated with all the data plots created as a result of user activity: the *TIC-chromatogram* window displays the TIC chromatograms of all the loaded data files and any XIC chromatogram created by the user; the *Mass-spectrum* window displays all the mass spectra obtained as a result of mass data combinations. If IM–MS data are loaded, the *Color-map* window displays the mass spectra *vs* drift times color map and the *Drift-spectrum* window displays all the drift spectra resulting from mass data integrations. Upon loading of a mass data file, a TIC chromatogram is automatically computed and, in case of IM–MS data, a color map is computed also (Figures 1 and S3).

## Rendering of mass datasets from different vendors

Successful combination of mass spectral data from different vendor data files is a challenge *per se*, because these data files may contain m/z data structured in distinct ways. For example, m/z data heterogeneity in a given dataset might be due to the mass spectra having 1) a variable number of data points, 2) varying origin m/z values, or 3) different intervals between neighbouring m/z values (that is, varying m/z steps). This mass data heterogeneity might only be problematic when performing mass data integrations to a mass spectrum (neither to a XIC chromatogram nor to a drift spectrum). Code was therefore developed to let the user tweak the configuration of the mass data combinations to a mass spectrum. The user is presented with a number of options involving (or not) explicit binning of the m/z data, and with preset values determined on the basis of a statistical analysis of the whole mass dataset (Figure S2). In the vast majority of all cases, the binning configuration automatically determined by statistical analysis of the whole mass spectral dataset yields perfectly fine mass spectra. However, tweaking the m/z integration paramaters may prove useful to refine the obtained mass spectrum in some corner cases or to accelerate the combination process when

datasets are very large (by increasing the size of the bins to an acceptable level).

## Chaining mass data integrations

A typical mass data visualization session starts within the TIC chromatogram that is computed upon loading of a mass data file (and the color map, in case of IM–MS data). Data integrations allow probing the data in all their depth, as illustrated in compact form in Figure 1 and, in a form more closely representing the user interactions with the graphical interface, in Figures S3 to S6.

After an IM–MS data file has been loaded (size-exclusion chromatography of bovine alcohol dehydrogenase (ADH) in native conditions), the initial plots are shown in the top panel of Figure 1 and Figure S3. The mass spectral data integration to a mass spectrum, starting either from the TIC chromatogram or from the color map, is performed by combination of all the mass spectra that have been acquired in the selected retention time region (or the drift time / mz region, in case of IM–MS data). All combinations to a mass spectrum are configured according to Figure S2.

The mass spectrum that is obtained shows the various oligomerization states of the alcohol dehydrogenase and their varying ionization levels (upper middle panel of Figure 1 and Figure S4). In order to determine the drift time of the various oligomerization states, integrations to a drift spectrum are performed over each m/z region corresponding to an oligomerization state of interest. Each integration produces a new drift spectrum (lower middle panel of Figure 1 and Figure S5).

The three drift spectra show that, whatever the charge states of the ions, the monomer (A), dimer (B) and tetramer (C and D) have significantly increasing drift times: dt (A) $<$ dt (B) $<$ dt (C, D): the higher the oligomerization state, the greater the drift time in the ion mobility cell. To determine at what retention time each of the various ions have eluted, integrations to a XIC chromatogram over the various regions of interest in the drift spectra are performed (bottom panel of Figure 1 and Figure S6). The results show that the tetramer

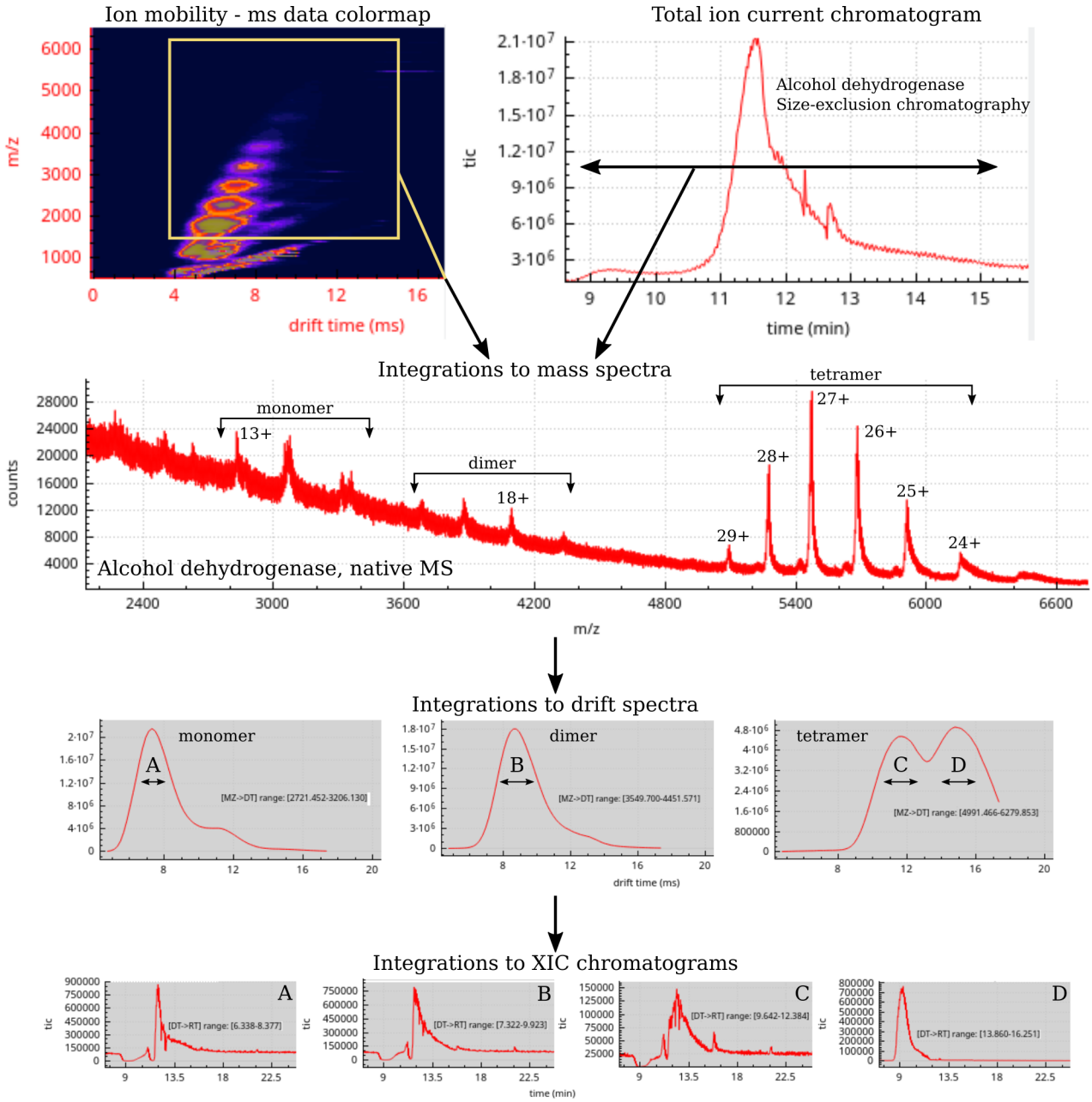



Figure 1: Mass data integrations can be performed in sequence. Each time a new integration is performed, the generated data subset is displayed in a new plot widget that may serve as a new integration starting point. This process allows one to dig into the depths of mass spectral data in a reasoned sequential manner.

elutes first and that the dimer elutes over a longer retention time range than the monomer, which is expected because this is a size-exclusion chromatography.

The various mass peaks belonging to the charge state envelopes of the monomer and oligomers might then be individually integrated to drift spectra. In this case (not illustrated), the resulting plots would show that the higher the charge of the tetramer (charge range $24^+$–$29^+$), the smaller the drift time, as would normally be expected. Back-integration of drift peaks C and D to a mass spectrum would show that drift peak C is not related to the protein but to another analyte. All the mass spectral integrations detailed in this example show that the capability to perform chained integrations in any direction allows one to very finely scrutinize mass spectral data in their full depth.

It is worthwhile noting that the original dataset loaded into the computer's memory did not change along the various integrations described above. The data consistency is preserved throughout successive integrations *via* a mechanism that records internally all the integrations that have led to a specific data plot (these internal integration details, called "history", are described in detail in the user manual provided as Supplementary Information). When that data plot is used as a starting point for a new integration, that history is transmitted to the mass data integration engine so that it performs the new integration only on the subset of the original dataset that accounts for all the history steps that occurred prior to this last integration. The fact that the original dataset is always available for new integrations permits proper comparison of mass spectral features. This feature sets *mineXpert* apart from the Driftscope software that is shipped with the Waters Synapt instruments.

## Mass peak deconvolutions

When mining data or when assessing the quality of a mass spectral acquisition, a procedure that is routinely performed is the determination of the molecular mass of analytes detected in a mass spectrum. All the mass spectrum-displaying widgets feature the possibility to deconvolute mass peaks according to the two commonly used methods: isotopic cluster-

based and charge state-based deconvolutions (Figure S7).

## Comparisons of mass spectral feature intensities

Analyzing mass spectral data most often requires comparing the intensity of features from different mass datasets. The "single TIC intensity value" integration computes the TIC intensity of any data feature found in the dataset, be it an isotopic cluster, a mass or drift peak, a chromatogram peak, or a color map region. Whatever the feature, the TIC intensity is computed by going back to the original mass spectral dataset first restricted to the data subset represented as the feature at hand (see above for the history concept).

For example, to compute the intensity of the signal encompassing the m/z region [5017–6234] of the mass spectrum in Figure 1 ([$24^+$–$29^+$] charge levels of tetrameric ADH), the user would right-click drag the mouse cursor over that m/z region while pressing the 'I' keyboard key. The single TIC intensity value of the data feature is shown in the status bar of the *Mass-spectrum* window in Figure S4. The interesting feature here, is that if the same process were applied to the whole drift time range of the third drift spectrum (tetramer, Figure S5) corresponding to that same m/z region, then the single TIC intensity value would be the same as that obtained above. This is because the drift spectrum plot is equivalent, from a mass spectral data standpoint, to the mass spectrum plot in region [5017–6234], as both are different representations of the same subset of the original dataset.

## Isotopic cluster calculations

Isotopic cluster calculations are often required in order to verify that a given acquired ion signal actually corresponds to a given ion of interest. This kind of pattern-matching work is typically used in top-down proteomics or in studies involving heavy isotopes. The IsoSpec project[12] implements a powerful algorithm in the form of a high performance library published under a Free Software license. To be used in *mineXpert*, the IsoSpec software was packaged for Debian and also built in the MS Windows and MacOSX ports of *mineXpert*.

IsoSpec is not easily configurable and *mineXpert* provides a user-friendly interface to perform the various kinds of configurations and user interactions needed to produce a mass spectrum describing the isotopic cluster (the steps are described in great detail in the user manual provided as Supplementary Information).

## Slicing of very large mass spectral datasets

When performing hyphenated techniques mass spectrometry, in particular with size exclusion chromatography, the mass spectral datasets are very large but the useful mass data might be much smaller, as in the example used above to illustrate the data integrations (see the TIC chromatogram in Figure 1, where useful data are located between min 9 and min 13 over a 30 min-long run). In order to reduce the size of huge datasets, code was developed to slice the original dataset into smaller chunks having an exact data subset copy of the original data (Figure S8).

## Data processing automation using JavaScript

In *mineXpert*, data processing automation was designed in such a manner that JavaScript statements can be issued either in a scripting console window or from a script file that might be read at program launch or from within the scripting environment. Figure 2 displays the main scripting environment window.

The scripting environment allows one to control the behaviour of all the windows containing plot widgets and the plot widgets themselves. Data from the plot widgets can be extracted to JavaScript objects that can then be used in the scripting environment according to the standard JavaScript syntax. Any kind of data can be exported to text files and plots can be exported to graphics data files. A history feature allows one to easily recall past commands (Figure S9) and a complete JavaScript developer reference documentation is available (Figure S10).

With JavaScript, *mineXpert* allows the user to conceive data processing and quality

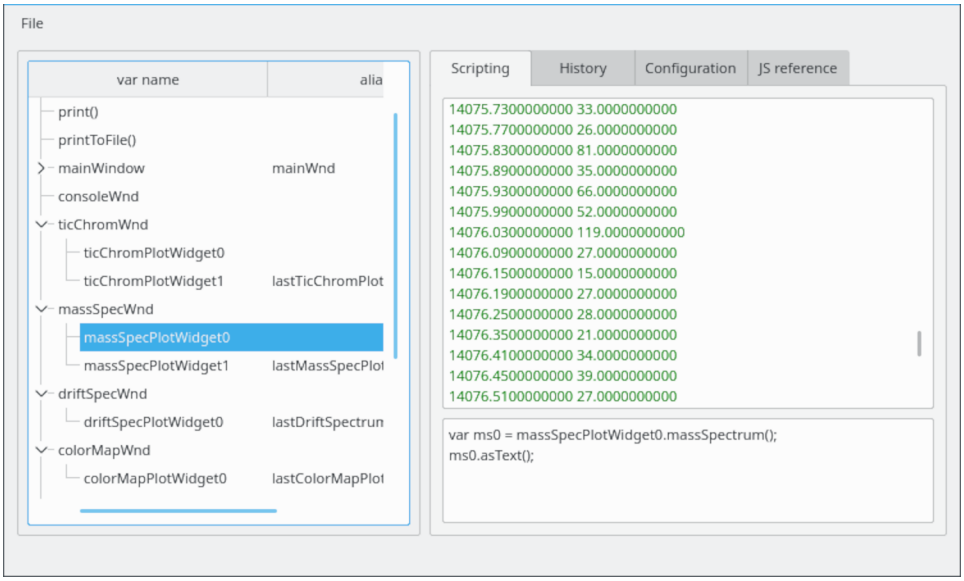



Figure 2: JavaScript-based automation of data processing tasks. The scripting environment lists, on the left hand side of the window, all the objects that have been produced by the software all along its use. Data plot widgets can be controlled as if by mouse using specific functions that are detailed in the JS reference tab. In this example, a MassSpectrum object is created from the data currently displayed in the `massSpecPlotWidget0` widget and then output in a textual form. Combinations or subtractions can be performed using pure JavaScript objects.

.

assessment workflows that can run unattended and thus be totally automated. A typical use case is when calibrating an IM–MS instrument's mobility cell. At that step, the mass analyzer has already undergone calibration, making mass measurements highly reliable, thus ensuring that the script code carries out flawlessly all the steps involved in the mobility cell calibration. Another interesting use case of scripting is when scientists who share the same datasets want to share data mining procedures to discuss the results. In this case, sharing a simple script allows anyone to replay in an identical manner the data mining steps (integrations, subtractions, zooming, panning, data export) that the script author performed in the first place. Finally, in quality control procedures for proteomics or metabolomics experiments, where internal standards are used in a series of mass spectrometric acquisitions, automatic control of these standards can be achieved using script-based automation.

## Data verification, mining and reporting

Very often, while mining data or performing quality assessments, recording of the findings is either desired or required. A recording system was developed to allow a flexible reporting of any finding, either to the clipboard, to a file, to the program console or to any combination of the three. The recording of the findings works exactly the same way in all the data plot widgets, TIC or XIC chromatogram, mass or drift spectrum. The kind of data that are recorded are, however, different: in chromatograms, the recorded data are typically intensities *versus* retention times; in mass spectra, intensities *versus* m/z values; in drift spectra, intensities *versus* drift times. To cope with that diversity of data, a specific syntax was devised such that the user may flexibly report any kind of data (Figure S11). While mining mass data, the user may record interesting findings by hitting the space bar.

# Conclusions

The *mineXpert* software program described in this report will be of use to a wide mass spectrometry audience because it allows mass spectrometry facility operators and users to inspect, validate and mine mass spectrometry datasets without the need of the closed-source software shipped with the instruments. *mineXpert* features the ordinary functionalities found in proprietary software systems and implements original mass data viewing paradigms and reporting of data mining discoveries. The software capability to handle ion mobility–mass spectrometry data will be of interest to the ever greater community of mass spectrometrists who implement this new technique. The isotopic cluster calculation functionality will be of interest to the top-down proteomics scientists. The automation of all the data loading and integration processes is of great value to mass spectrometry facility operators and users willing to set up unattended processing workflows.

# Acknowledgement

The author thanks Dr Valérie Gabelica for providing IM-MS datasets that have served to test the software with Agilent instrument-produced files; Prof. Guillaume van der Rest and Dr Frédéric Halgand, for many enjoyable and interesting discussions and for the Waters Synapt2-originating dataset used to illustrate *mineXpert* in this report; the PAPPSO team at the GQE-Moulon laboratory, for fruitful discussions on proteomics-oriented mass spectrometry data acquisition and analysis; Dr Olivier C. Martin for the critical reading of the manuscript.

# Supporting Information Available

The following supporting information is available free of charge at ACS website http://pubs.acs.org:

- supplementary-information.pdf: detailed figures and explanations of the *mineXpert* program features.

- minexpert-doc.pdf: user manual of the *mineXpert* software program.

# References


(1) Röst, H. L. et al. OpenMS: A flexible open-source software platform for mass spectrometry data analysis. *Nat Methods* **2016**, *13*, 741 – 748.

(2) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: Open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534 – 2536.

(3) Langella, O.; Valot, B.; Balliau, T.; Blein-Nicolas, M.; Bonhomme, L.; Zivy, M. X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *J. Proteome Res.* **2017**, *16*, 494 – 503.

(4) Valot, B.; Langella, O.; Nano, E.; Zivy, M. MassChroQ: A versatile tool for mass spectrometry quantification. *Proteomics* **2011**, *11*, 3572 – 3577.

(5) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11*, 395.

(6) Avtonomov, D. M.; Raskind, A.; Nesvizhskii, A. I. BatMass: A Java Software Platform for LC-MS Data Visualization in Proteomics and Metabolomics. *J. Proteome Res.* **2016**, *15*, 2500 – 2509.

(7) Strohalm, M.; Kavan, D.; Novak, P.; Volny, M.; Havlicek, V. MMass 3: A cross-platform software environment for precise analysis of mass spectrometric data. *Anal. Chem.* **2010**, *82*, 4648 – 4651.

(8) Sivalingam, G. N.; Yan, J.; Sahota, H.; Thalassinos, K. Amphitrite: A program for

processing travelling wave ion mobility mass spectrometry data. *Int. J. Mass Spectrom.* **2013**, *345 - 347*, 54 – 62.

(9) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, *26*, 966 – 968.

(10) Rusconi, F.; Belghazi, M. Desktop prediction/analysis of mass spectrometric data in proteomic projects by using massXpert. *Bioinformatics* **2002**, *18*, 644 – 645.

(11) Rusconi, F. MassXpert 2: A cross-platform software environment for polymer chemistry modelling and simulation/analysis of mass spectrometric data. *Bioinformatics* **2009**, *25*, 2741 – 2742.

(12) Lacki, M. K.; Startek, M.; Valkenborg, D.; Gambin, A. IsoSpec: Hyperfast Fine Structure Calculator. *Anal. Chem.* **2017**, *89*, 3272 – 3277.

(13) Gabelica, V.; Livet, S.; Rosu, F. Optimizing Native Ion Mobility Q-TOF in Helium and Nitrogen for Very Fragile Noncovalent Structures. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 2189 – 2198.

(14) Deutsch, E. W. et al. Proteomics Standards Initiative: Fifteen Years of Progress and Future Work. *J. Proteome Res.* **2017**, *16*, 4288 – 4298.

(15) Deutsch, E. W. Mass spectrometer output file format mzML. *Methods Mol Biol* **2010**, *604*, 319 – 331.

(16) Holman, J. D.; Tabb, D. L.; Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Curr Protoc Bioinformatics* **2014**, *46*, 13 – 24 – 1 – 9.

(17) Van der Rest, G.; Halgand, F. Size Exclusion Chromatography-Ion Mobility-Mass Spectrometry Coupling: A Step Toward Structural Biology. *J. Am. Soc. Mass. Spectrom.* **2017**, *28*, 2519 – 2522.

# Graphical TOC Entry

