PeakQC: A Software Tool for Omics-Agnostic Automated Quality Control of Mass Spectrometry Data

Andrea Harrison[1], Josie G. Eder[2], Priscila Lalli[2], Nathalie Munoz[1,4], Yuqian Gao[2,4], Chaevien S. Clendinen[1], Daniel J. Orton[2], Xueyun Zheng[2], Sarah M. Williams[1], Sneha P. Couvillion[2], Rosalie K. Chu[1], Vimal K. Balasubramanian[1], Arunima Bhattacharjee[1], Christopher R. Anderton[1], Kyle R. Pomraning[3,4], Kristin E. Burnum-Johnson[1,4], Tao Liu[2], Jennifer E. Kyle[2], and Aivett Bilbao[1,4]*

[1] Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, 99352, USA
[2] Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, 99352, USA
[3] Energy Processes & Materials Division, Pacific Northwest National Laboratory, Richland, WA, 99352, USA
[4] US Department of Energy Agile BioFoundry, Emeryville, CA, 94608, USA

* **Correspondence**: Aivett.Bilbao@pnnl.gov

Abstract

Mass spectrometry is broadly employed to study complex molecular mechanisms in various omics biological and environmental research, including proteomics, metabolomics, and lipidomics. As study cohorts grow larger and more complex with dozens to hundreds of samples, the need for robust quality control (QC) measures through automated software tools becomes paramount to ensure the integrity, high quality, and validity of scientific conclusions from downstream analyses and minimize waste of resources. Since existing QC tools are mostly dedicated to proteomics, automated solutions supporting metabolomics are needed. To address this need, we developed the software PeakQC, a tool for automated QC of MS data which is independent of omics molecular types (i.e., omics-agnostic). It allows automated extraction and inspection of peak metrics of precursor ions (e.g., errors in mass, retention time, arrival time) and supports various instrumentations and acquisition types, from infusion experiments or using liquid chromatography and/or ion mobility spectrometry front-end separations, and with/without fragmentation spectra from data-dependent or independent acquisition analyses. Diagnostic plots for fragmentation spectra are also generated. Here, we describe and illustrate PeakQC's functionalities using different representative datasets, demonstrating its utility as a valuable tool for enhancing the quality and reliability of omics mass spectrometry analyses.

Keywords

liquid chromatography, ion mobility spectrometry, data-dependent acquisition, data-independent acquisition
data quality control
proteomics, metabolomics, lipidomics
algorithm, computational tool, data processing

Introduction

Mass spectrometry (MS), and its versatile coupling with various analytical separation techniques, enables analysis of a wide range of different molecular types, extending from small compounds to large proteins. Its applications span many science areas, including clinical, pharmaceutical, forensic, material, environmental, and food analysis. During a typical MS-based omics study (e.g., proteomics, metabolomics, lipidomics), different sources of variability can arise at any stage of the experimental workflow, from sample preparation to chromatography separation, MS detection, and data processing[1, 2]. These sources of variability can significantly impact the performance and results of the experiment, potentially leading to inaccurate interpretations. Therefore, the implementation of quality control (QC) strategies during sample MS analysis and the use of computational tools for monitoring and assessing the instrument performance is crucial to ensure the high quality, reproducibility, and accuracy of the data generated and utilized in downstream analyses to produce high-confidence results.

Particularly in large-scale untargeted MS studies, early detection of instrument's performance issues is necessary to avoid waste of resources, such as loss of precious samples, instrument time, computer storage space (i.e., data incorrectly acquired and thus unusable), and data processing computing resources and labor (i.e., processing of flawed data and expert human curation of incorrect results). Numerous tools have been reported to facilitate the QC of MS data (see examples in Table 1). Most of these software tools are designed for proteomics and utilize the output of search engines as a common method to perform QC based on the detection of expected peptides and evaluation of identifications (IDs). This method works well for proteomics, because many available identification tools can be run in an automated fashion and produce reliable results. However, they do little to pinpoint the specific causes of performance deterioration, because a decrease in the total number of spectral matches could be caused by a variety of reasons, such as deteriorated chromatography performance, decreased MS sensitivity, or poor mass calibration[3]. Furthermore, many identification tools for metabolomics and lipidomics require manual expert curation, where errors are usually detected after the data processing and downstream analyses are performed. In many cases this could be too late to reproducibly re-run the analysis of samples to replace the compromised data.

While there are active community efforts to establish and promote best practices for QC in untargeted metabolomics[4-6], the number of available software tools for QC in metabolomics is limited. Traditional QC approaches often involve manual inspection of spectra, which is time-consuming, subjective, and prone to human error. Operators frequently need to inspect the compromised data using the instrument vendor’s software. However, these tools typically do not support easy and fast comparison across multiple samples beyond visualization, and in cases where the instrument is failing or a sample is damaged, the operators need to open and inspect the data from one sample at a time to identify and fix the issue before continuing the analysis.

Most tools in Table 1 incorporate variables from precursor spectra (MS1) and tandem spectra (MS/MS or MS2) to calculate metrics to assess QC and generate visualization plots. However,

existing tools have limitations: they are tailored to a single omics type, they can be difficult to deploy, they may lack active maintenance, do not support multiple MS acquisition schemes, and/or do not support novel hybrid MS platforms such as ion mobility spectrometry. Moreover, with a higher number of multiomics studies integrating proteomics, metabolomics, and lipidomics data, the need for omics-agnostic QC tools is becoming increasingly evident. Therefore, a pressing demand exists for automated QC solutions capable of handling diverse data types and scaling with the growing volume of data generated in contemporary MS-based omics experiments.

To address these challenges, we developed PeakQC, a novel software tool designed for omics-agnostic automated QC of MS data. It performs automated extraction of QC metrics and supports various instrumentations and acquisition schemes, from infusion experiments, to liquid chromatography (LC), and ion mobility spectrometry (IMS)[7] front-end separations, and with/without fragmentation spectra from data-dependent acquisition (DDA) or data-independent acquisition (DIA) modes[8]. We illustrate how PeakQC applies omics-agnostic algorithmic strategies and leverages machine learning techniques to provide a robust framework for rapid, accurate, and unbiased assessment of MS data quality.

1 Table 1. Software tools freely available for quality control in untargeted MS.
2 PCA: Principal component analysis; TIC: total ion chromatogram; XIC: extracted ion chromatogram; RT: retention time; AT: arrival time; S/N: signal to noise; FWHM:
3 full width at half maximum.

Software, reference	Omics type, main algorithms	MS1 main metric variables	MS2 main metric variables	Input formats	Prog. language	Description comments
PeakQC, (this work)	Omics-agnostic, PCA, Isolation Forest, QC metrics calculation	Count spectra, TIC, XIC, RT error, AT error, Mass error	Count spectra, TIC	Thermo, Agilent, mzML (through MZA)	Python	Applies an automated QC analysis workflow based on ions and independent of molecular identification tools. Ions can be user-specified or automatically detected (auto-tracked) to extract metrics and generate reports. Supports LC, IMS, DDA, DIA, and infusion analyses. Desktop application.
Rapid QC-MS Sandhu et al. 2024[9]	Metabolomics QC metrics calculation	Features, RT error, Mass error	Spectral similarity	Various vendors (through mzML MSConvert), Reports from other tools (MS-DIAL)	Python	Provides an automated, interactive dashboard for assessing LC-MS/MS QC results spanning multiple performance dimensions and applying thresholds for QC metrics to pass or fail.
DO-MS, Wallmann et al. 2023[10]	Proteomics QC metrics calculation	Accumulation time, TIC, S/N, Features, IDs	Accumulation time, TIC, Count spectra	Thermo, mzML, Reports from other tools (Dinosaur, DIA-NN)	Python, R	Performs a data-driven optimization of the MS method and quality control of DIA experiments. Allows for interactive data display and generation of portable html reports through an interactive R Shiny app.
MetaPro, Ann et al. 2023[11]	Metabolomics QC metrics calculation	XIC, RT		Aird	Java, Java Script	Supports a semi-targeted analysis workflow for DDA and various functions for fast QC inspection and spectral library curation with easy-to-use interfaces. Web-based application.
LipidSpace, Kopczynski et al. 2023[12]	Lipidomics QC metrics calculation	Abundance CV, IDs		Reports from other tools (IDs and quantitation tables from lipid search engines)	C++	Allows analyzing lipidomes by assessing their structural and quantitative differences and offers an interactive workflow that guides the user through QC steps with statistic figures and visualizations such as dendrograms.
RawBeans, Morgenstern et al. 2021[13]	Proteomics QC metrics calculation	Injection time, TIC, Intensity, Mass error	Injection time, Count spectra, TIC	Thermo, mzML	Python	Generates HTML-based reports with figures for user inspection that include the key parameters needed to monitor the LC-MS system performance.
QC-ART, Stanfill et al. 2018[14]	Proteomics, Dynamic linear model, QC metrics calculation	Count spectra, TIC, IDs	Count spectra, Reporter ions (if iTRAQ data)	Reports from other tools (MASIC, MSGF+)	R	Uses a dynamic linear model to flag anomalies or potential issues with instrument performance or sample quality changes over time. Identifies local and global deviations in data quality.
Panorama AutoQC, Bereman et al. 2016[3]	Proteomics QC metrics calculation	XIC, Peak area, FWHM, RT error, Mass error, Isotopic dot product		Agilent, Bruker, Thermo, Waters, Shimadzu, Sciex	C#	Provides capabilities to store and visualize QC data in a longitudinal fashion. A stand-alone program that can be installed and run on instrument control computers to automatically import QC data files into a Skyline document and upload it to the Panorama Server.
iMonDB, Bittremieux et al. 2015[1]	Proteomics QC metrics calculation			Thermo	Java	Allows to automatically extract, store, and manage instrument parameters (e.g., ESI capillary temperature and voltage) from raw-data objects into a highly efficient database structure, enabling monitoring over time.
LLRC, Amidan et al. 2014[15]	Proteomics Lasso Logistic Regression Classifier (LLRC), QC metrics calculation	Count spectra, IDs, XIC, Peak area, FWHM, RT error, Mass error	Count spectra	Reports from other tools (SMAQC, Quameter, MSGF+)	R	A composite classifier implemented as an R package that detects compromised data with high sensitivity while maintaining high specificity.
Metriculator, Taylor et al. 2013[16]	Proteomics QC metrics calculation	Count spectra, XIC, FWHM, S/N, IDs	Count spectra, Intensity, S/N	Thermo, NIST metrics (MSQC)	Ruby	Generates interactive comparison plots of metrics for determination of outliers and trends in the datasets, together with relevant statistical comparisons.
SIMPATIQCO, Pichler et al. 2012[17]	Proteomics QC metrics calculation	Count spectra, TIC, XIC, RT FWHM,	Count spectra, TIC	Thermo, Sciex. Reports from other tools (MASCOT)	PHP	Assists the longitudinal monitoring of QC metrics and LC-MS system performance. Results are stored in a database and can be displayed via a web browser. For each QC metric the software learns the range

		IDs				indicating adequate system performance using robust statistics.

Methods

Software design and implementation

PeakQC was implemented in Python. A minimalist design was applied for the graphical user interface (GUI) to provide an easy-to-maintain, and simple, but user-friendly selection of raw data files (i.e., MS runs) for unattended batch processing. Raw MS data is accessed directly in instrument format using MZA[18] (current supported formats: Agilent .d, Thermo .raw, Bruker .d, and mzML). Python packages used include tkinter, h5py, multiprocessing, numpy, pandas, scikit-learn and matplotlib. A set of metrics that are applicable across different types of molecules are automatically calculated. Metrics include errors in mass, retention time (RT), and arrival time (AT, also referred to as drift time for drift-tube type of ion mobility systems), and spectra data metrics such as number of spectra (count spectra) and total ion chromatogram (TIC), which are calculated, analyzed, and visualized across all MS runs within the groups in the sample analysis batch. An overview of PeakQC software architecture and algorithm is shown in Figure 1. The PeakQC software executable, user guide, source code, and example data are available at https://github.com/pnnl/IonToolPack.

Algorithm

Given a list of raw data files acquired from samples (i.e., MS runs), the tool performs the following steps automatically to: 1) Convert the raw MS data to MZA format. The converter is included along with the PeakQC executable, and the conversion is automatic and transparent to the user; 2) Generate a time-vs-*m/z* image of each MS run with a very fast algorithm; 3) Perform a principal component analysis (PCA) of the MS runs based on the images. Samples are colored by their respective group, which is either automatically or user assigned. The PCA plot remains open and is interactive (e.g., zoom in and out, move) and the initial PCA figure is also automatically saved as a PDF; 4) Find ions in the MS1 data for auto-tracking by automatically detecting the topmost abundant ions that are more frequent per sample group; 5) Extract peak metrics for both auto-tracked ions and optionally user-specified ions (i.e., theoretical or reference values); 6) Generate reports (formatted in comma-separated values, CSV) and figures (JPG, PDF) to visualize and track different ion metrics for the auto-tracked ions found and for the user-specified ions. The user can find and navigate the generated result files organized in output folders; 7) Perform outlier detection per metric and sample group employing the Isolation Forest method; 8) Generate error heatmaps to display defective MS runs and corresponding metrics with issues.

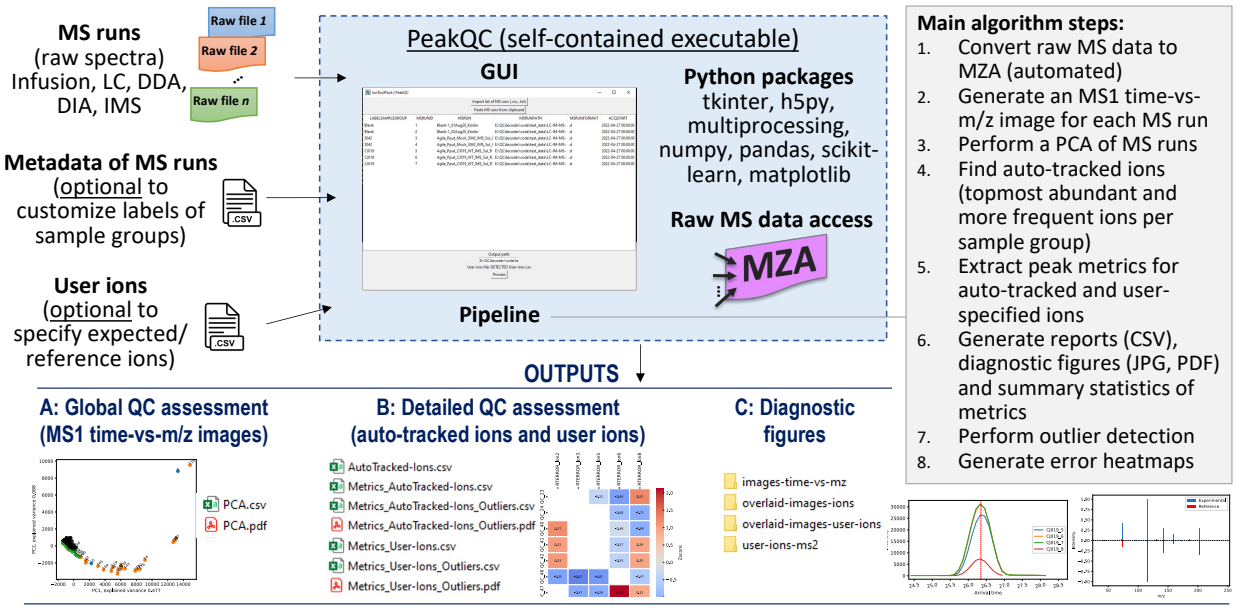

Figure 1. **Overview of PeakQC software architecture and algorithm**. A simple and user-friendly GUI allows easy selection of input files and launching the automated processing pipeline, which executes the main algorithm steps shown on the right (gray box). The conversion of the raw data to MZA format is completely automated and requires no user intervention. MZA facilitates software development and tool deployment, and it is well-suited for handling raw data from various types of instrumentation beyond LC-MS, including IMS and DIA. PeakQC is packaged as a self-contained executable, which includes all necessary dependencies (i.e., MZA and various Python packages). This executable requires no installation; once downloaded, the user can directly launch the tool by double-clicking it. During processing, several output files organized in folders are generated to provide both global and detailed QC assessments and diagnostic figures. **A:** The PCA provides a first QC assessment to quickly spot compromised MS runs reflected by global similarity patterns in the MS1 data. **B:** Auto-tracked ions are automatically detected and the values from the first MS run are taken as reference to calculate error metrics. User ions can be optionally specified by the user with expected values. Extracted peak metrics allow detailed evaluation reflected by local differences per dimension and per sample group (e.g., errors in mass, retention time, arrival time). The heatmap only display the defective MS runs and corresponding defective metrics, i.e., the larger the heatmap the more issues detected: more rows indicate more defective runs, and more columns indicate more defective ions and metrics. **C:** Diagnostic figures are generated with overlaid extracted peak signals (e.g., XICs) and MS2 mirror plots, which the user can look at to further investigate issues.

Datasets

Various datasets were used for tool development and evaluation. The metabolomics and lipidomics LC-MS datasets were collected as part of the analyses performed for the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC, https://proteomics.cancer.gov/programs/cptac) with a 40-min C18 LC separation for lipidomics[19], and a 10-min HILIC LC separation for metabolomics , and coupled to a Thermo

Lumos mass spectrometer operated with negative ESI (electrospray ionization) and in DDA mode for tandem MS. The metabolomics direct infusion MS dataset corresponds to a plant experiment from a modified version of a previously published platform[20] using a Thermo QExactive mass spectrometer operated with negative ESI. The metabolomics LC-IMS-MS and proteomics LC-MS datasets are part of a previous publication of an Agile BioFoundry (http://agilebiofoundry.org) synthetic biology study[21], with a 10-min HILIC LC separation for metabolomics, coupled to an Agilent 6560 Drift Tube IMS-QTOF mass spectrometer operated with negative ESI and in All Ions DIA mode, and the PNNL-PreProcessor used for data preprocessing[22], and a 120-min RP LC separation for proteomics, coupled to a Thermo QExactive mass spectrometer operated with positive ESI and DDA mode.

Results and discussion

PeakQC is a desktop and stand-alone tool that supports various MS acquisition types, including LC-MS, direct infusion MS, LC-IMS-MS, DDA, and DIA. The GUI allows the user to select input files and start the automated processing pipeline (Figure 1). The following sections describe and illustrate PeakQC's functionalities using different representative metabolomics, lipidomics, and proteomics datasets from various human health, environmental, and synthetic biology studies.

Fast QC assessment of batch of samples by PCA

The first functionality in PeakQC is a PCA of the batch of samples, where each MS run or sample in the batch is represented by a time-vs-*m/z* image generated from the MS1 data. In this PCA view, samples are colored by automatically assigned or user-specified groups (e.g., instrument blanks, sample preparation blanks, QC standards, pooled samples, sample treatments, etc.), providing a first QC assessment to quickly spot compromised MS runs, since runs of the same sample type should be properly clustered together and the clusters should reflect the sample groups. Examples of PCAs revealing defective MS runs are shown in Figure 2.

The time-vs-*m/z* images are generated from the raw spectra independently of peak detection or identification tools, which enables utilization of this PCA functionality for any MS analysis, independently of chromatography separations and omics type. In this way, the PCA view works even for MS runs that do not contain chromatographic separation (i.e., direct infusion analyses). In the case of IMS data, the IMS dimension is ignored for this step. In the time-vs-*m/z* image representation, the intensity is log-transformed, and thus distorted, and as Figure 2-A shows, individual peaks are not well resolved, but this representation is sufficient to capture global similarity patterns in the data.

PeakQC allows automatic or customized assignment of labels for sample groups. Automated assignment of sample groups is done by detecting the substrings ‘Blank’ and ‘QC’ in the file names (not case sensitive) and assigning ‘Sample’ otherwise. In addition, customized sample groups can be specified through an input file, like shown in Figure 2-B and C. The lipidomics LC-MS analysis shows two kinds of QC runs, QCs from pooled analyte samples and QCs from a standard lipid-extract sample. A recent review reported that the use of pooled QC samples for monitoring data quality has been relatively widely adopted by the metabolomics community [4].

The PCA functionality can be utilized by the operator near-real time during data acquisition of a batch of samples, by re-launching the QC analysis after a new MS run is completed, since monitoring during true real time while the instrument is acquiring data and before finishing the run would require communication with the instrument vendor software, which is rarely provided to external software. For evaluating the first MS runs in a new batch, a QC run from previous batches which are of good quality should be included in the analysis as reference, and in the case that multiple runs in the new batch are defective, then the reference QC run will be shown as outlier and any issue can be detected.

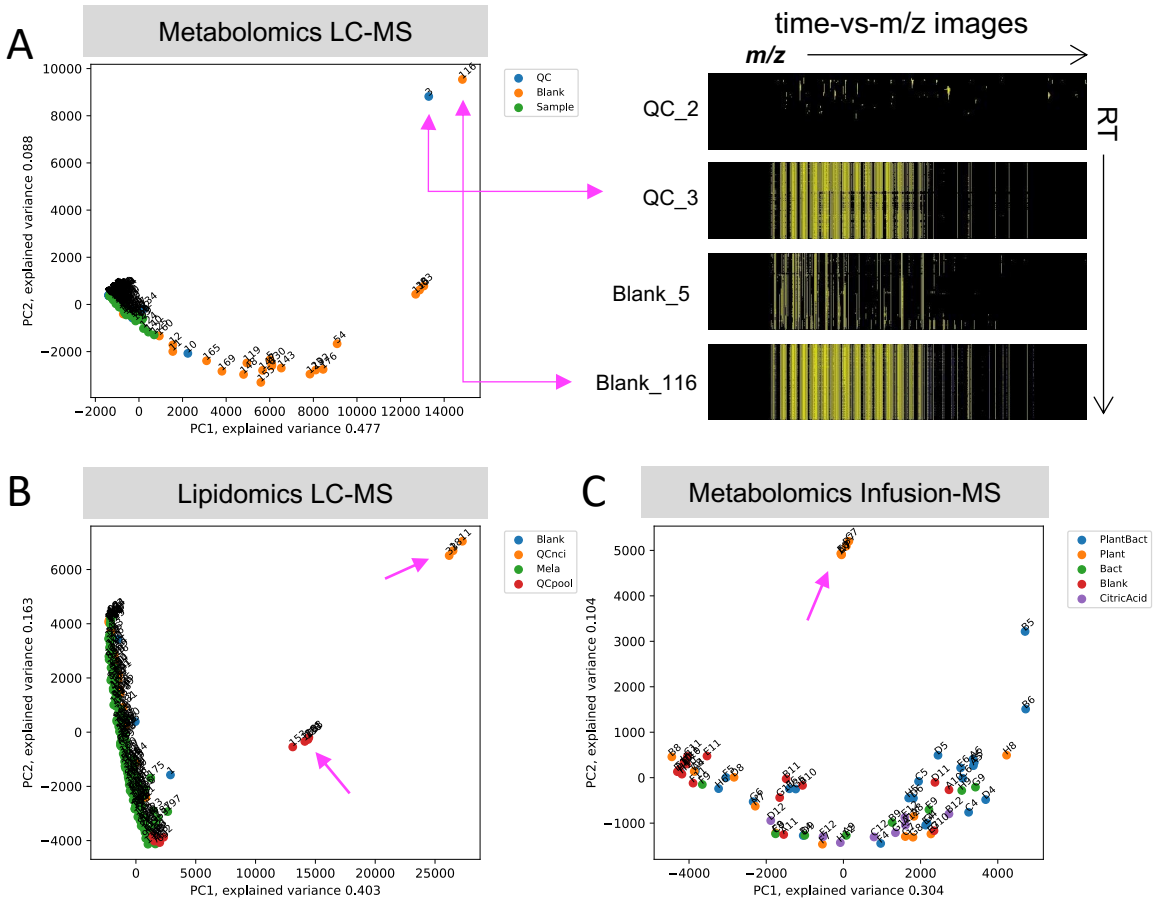

Figure 2. **Examples of PCA of batches of samples**. PeakQC automatically generates PCA scores and a plot as an interactive window and also saves them as CSV and PDF, respectively. Each dot is an MS run represented by a time-vs-m/z image generated from the MS1 spectra and colored according to its sample group. Outliers are pointed with pink arrows. **(A)** Metabolomics LC-MS analysis where 2 MS runs had no LC separation, which was disabled due to an issue in the LC system. The time-vs-m/z images on the right show pronounced vertical lines for the outlier MS runs QC_3 and Blank_116, which make them more similar to each other compared to other MS runs with enabled LC separation, for example QC_2 and Blank_5. **(B)** Lipidomics LC-MS analysis with sample evaporation issues in several QC runs of two types, QCs from pooled analyte samples (QCpool) and QCs from a standard lipid-extract sample (QCnci). The evaporation was

compensated by adding solvent, however, this resulted in excessive dilution and data with high background signals, which in turn produced the two separate QC clusters, one for each type of diluted QC. **(C)** Metabolomics infusion analysis of a plant experiment where samples were collected from at grid with several plants grown under different conditions and capturing samples spatially at different root lengths. The outliers correspond to samples collected from grid cells at the bottom, beyond the length of the root (i.e., no analyte). Panel A shows automated assignment of sample groups. Panel B and C show customized assignment of sample groups. Panel C shows customized assignment of sample IDs corresponding to the grid cells in the experiment.

Automated detection of anomalies in peak metrics

Following the PCA of the batch of samples, PeakQC finds the topmost abundant ions in the MS1 data which are more frequent per sample group and uses them as 'auto-tracked' ions. The number of auto-tracked ions can be changed in the PeakQC configuration file. Peak metrics are then extracted for both auto-tracked and user-specified ions (i.e., theoretical or reference values), and various reports (CSV) and figures (JPG and PDF) are generated. For auto-tracked ions, PeakQC takes the values from the first MS run as reference. Finally, an outlier detection analysis is performed for individual peak metrics for each sample group, and error heatmaps showing the detected outliers and corresponding metrics and MS runs are generated. PeakQC uses the Isolation Forest algorithm which detects anomalies using binary trees[23]; it has a linear time complexity and a low memory requirement, which works well with high-volume data. Figure 3 shows an example of detection of an outlier MS run in RT for the user-specified ion tryptophan from the metabolomics LC-MS analysis. The heatmap only display the defective MS runs and corresponding defective metrics, i.e., the larger the heatmap the more issues detected: more rows indicate more defective runs, and more columns indicate more defective ions and metrics. These heatmaps provide a focused view with important complementary information to the PCA, as the user can quickly see which MS runs and which metrics resulted in QC issues. To confirm the detected outliers, the user can look at the MS1 overlaid figures per ion and by sample group, as shown in the overlaid chromatograms in Figure 3-B.

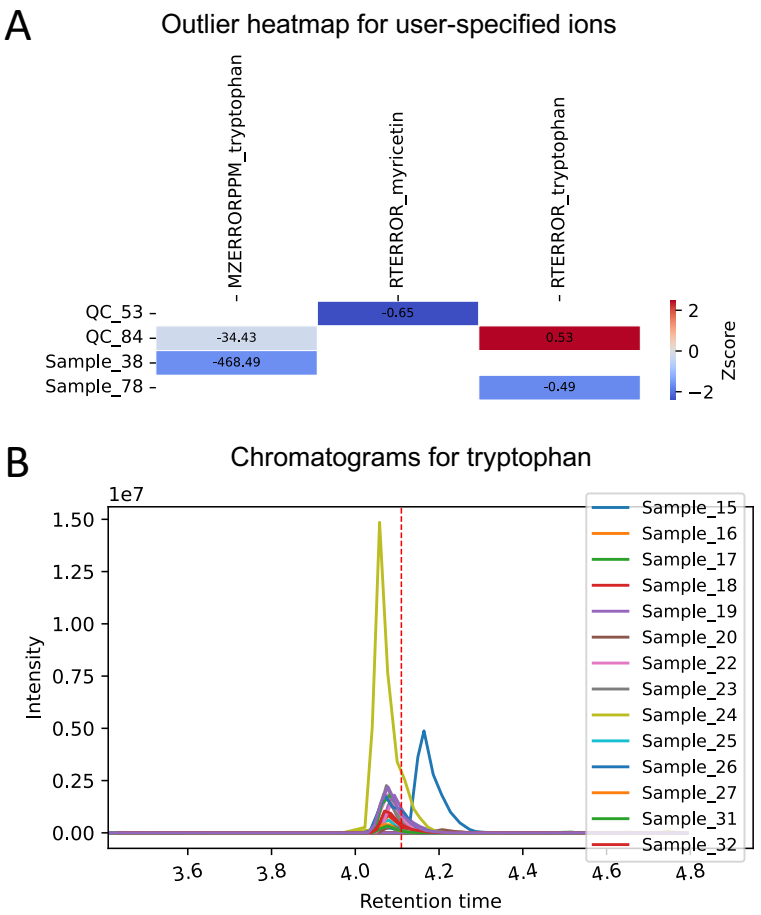

Figure 3. Metabolomics LC-MS analysis showing detection of an RT outlier. PeakQC perform outlier detection for individual peak metrics and for each sample group. (A) Heatmap of detected outliers and corresponding metrics and MS runs. The color of the cell indicates the Z-score calculated within the sample group and the text within each cell indicates the metric value (e.g., mass error in ppm and RT error in minutes). The Z-score indicates how many standard deviations above or below the mean error a peak metric is: a positive Z-score says the data point is above average and a negative Z-score says the data point is below average. (B) Chromatogram view for tryptophan and the 'Sample' group. The most intense peak (light green) detected 0.49 min earlier than the reference RT (dashed red line) corresponds to Sample_78 (name not shown in legend due to default truncation to accommodate display).

Figure 4 shows an example of detection of an outlier MS run with large mass errors in the metabolomics LC-IMS-MS analysis. While large mass errors are typically less common than other QC issues, such as compromised chromatography or decreased sensitivity, they can still occur due to various factors, for example, when the instrument was not regularly calibrated, or the sample was contaminated. For this example, a coefficient of the time-of-flight mass calibration was manually altered in a copy of an MS run to induce a large mass shift. As expected, the generated heatmaps for auto-tracked ions and user-specified ions show the MS run CJ019_8 as outlier with large *m/z* errors (Figure 4-B). To calculate errors for auto-tracked

ions, PeakQC takes the values from the first MS run as reference, therefore, including a previous QC run that is of good quality is required for the results from auto-tracked ions to be informative. The current software implementation does not extract AT for auto-tracked ions, however, mobilograms are extracted for user-specified ions if the AT is included in the input and the MS data contains IMS separation (example in Figure 4-C).

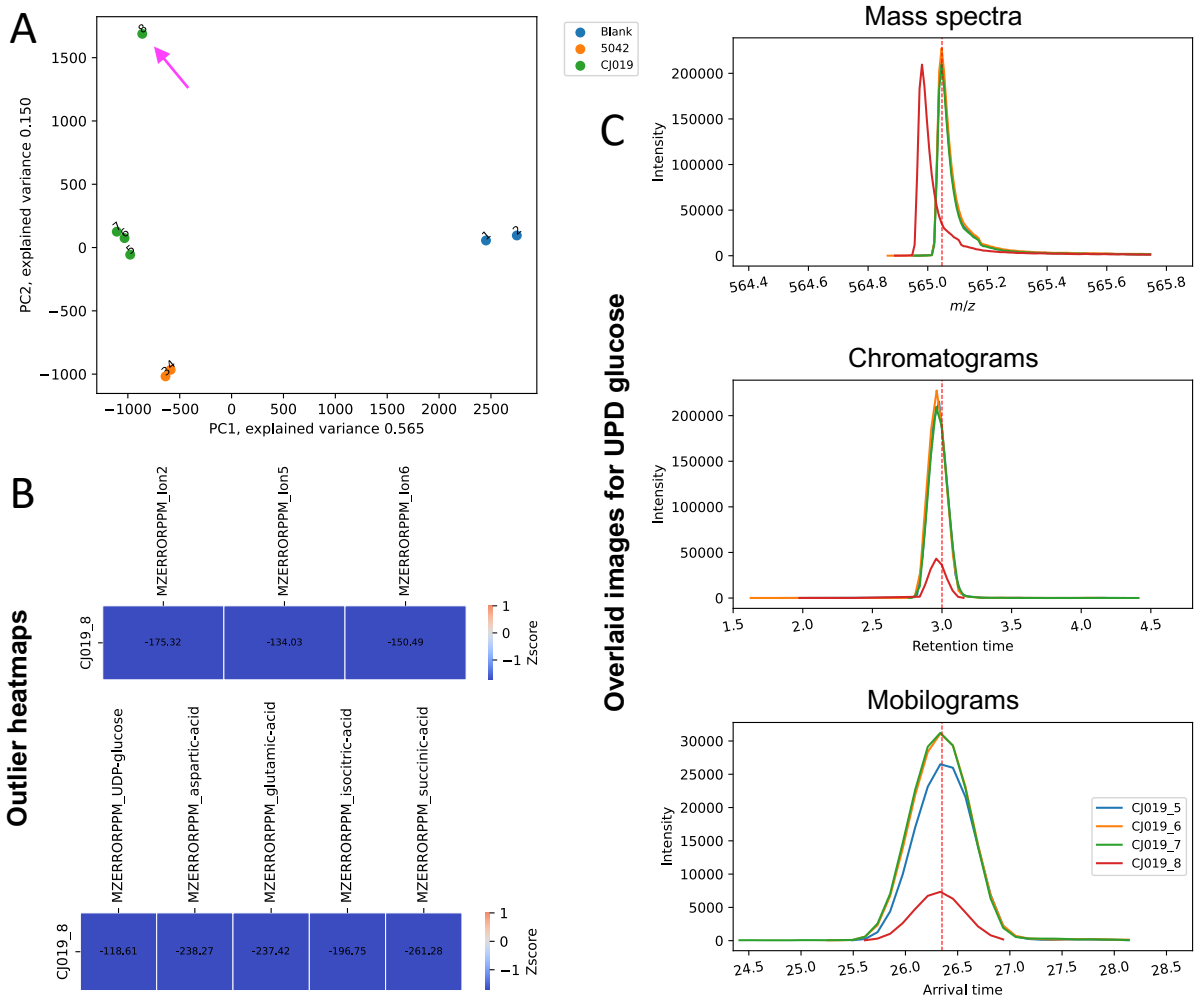

Figure 4. Metabolomics LC-IMS-MS analysis showing detection of an outlier MS run with mass error issues. **(A)** PCA indicating the MS run CJ019_8 with a pink arrow. **(B)** Outlier heatmaps of auto-tracked ions (top) and user-specified ions (bottom) reporting MS run CJ019_8 with large *m/z* errors. **(C)** Overlaid mass spectra, chromatograms, and mobilograms for the user-specified ion UDP glucose (*m/z* 565.05, RT 3.0 min, AT 26.35 ms) in sample group CJ019. PeakQC generates this kind of figure for each user-specified ion. Dashed red lines show the reference values which are indicated in the input file for the user-specified ions. The mass shift can be observed in the top panel for the red trace (CJ019_8). Mobilograms are extracted if the AT is specified in the input and the MS data contains ion mobility spectrometry separation.

Figure 5 shows an example of RT shifting issues in a proteomics LC-MS analysis. This issue was presumably due to an unstable injection flow. In addition to the outlier heatmaps, PeakQC generates error heatmaps showing ions in QC runs that contain metric values outside of user-specified tolerances. While some degree of RT shift is typical for longer proteomics separations and wider RT tolerance are used, the error heatmap of the QC runs allows easy and rapid inspection of expected RTs in the QC runs.

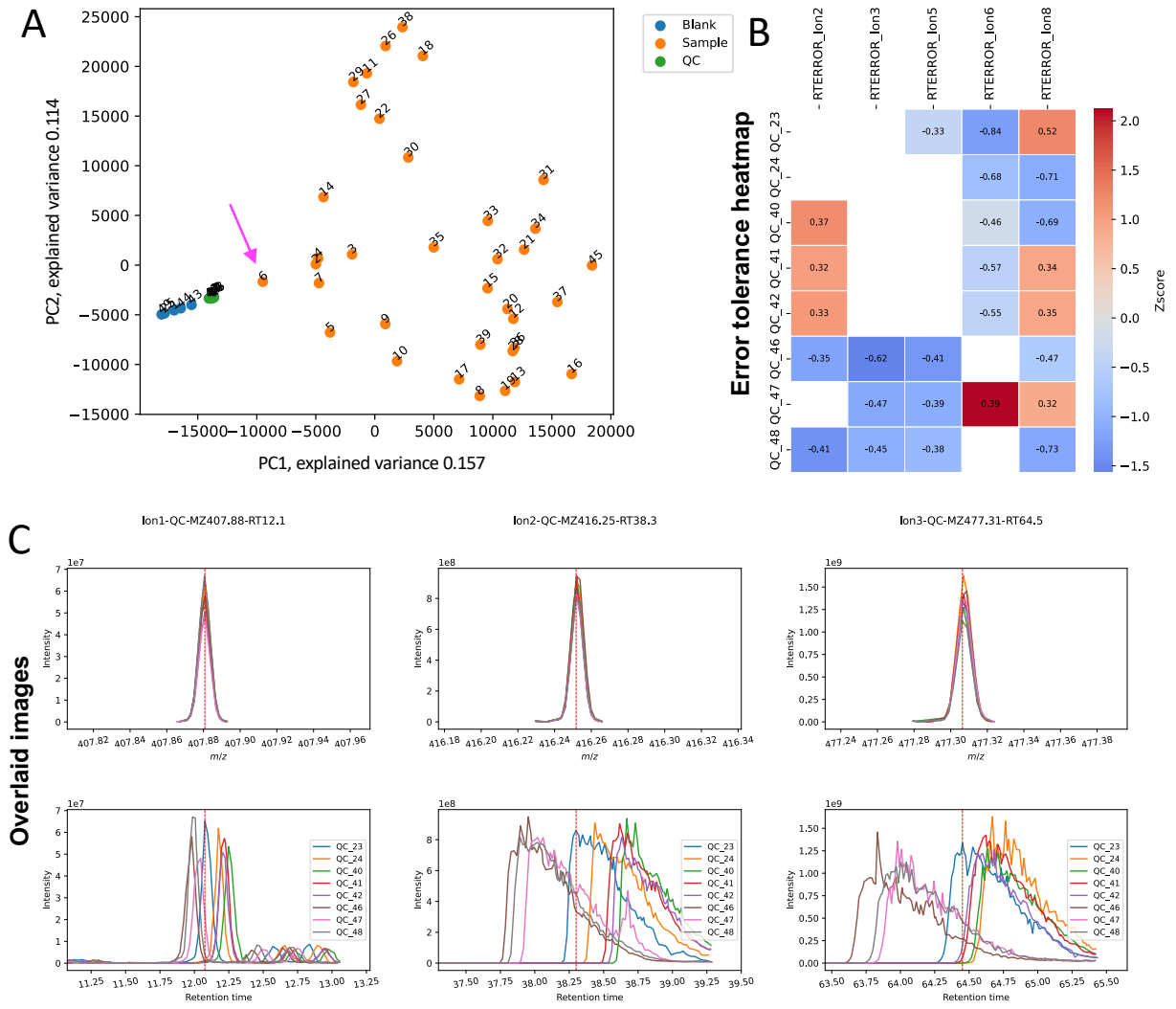

Figure 5. Proteomics LC-MS analysis with RT shifting issues. **(A)** PCA of the batch of samples. The MS run Sample_6 indicated with a pink arrow had the largest RT shift of its group (orange dots). **(B)** Error heatmaps of auto-tracked ions reporting the QC runs with RT errors outside the 0.3-min tolerance used. **(C)** Overlaid images showing the first 3 auto-tracked ions. Ion1 is not displayed in the heatmap because all the peaks are within tolerance. Ion2 and Ion3 show peaks outside the tolerance, for example, peaks from QC runs 46, 47 and 48 have earlier RT times, thus negative RT errors and negative Z-scores in the heatmap (blue cells).

Inspection of MS2 peak signals

In addition to the overlaid MS1 figures, PeakQC extracts signals from the MS2 data and generates diagnostic figures for each molecule and each MS run, for the user-specified ions, if their expected fragment ions and intensities are included in the input. Figure 6-A shows examples for LC-MS DDA spectra and Figure 6-B shows examples for LC-IMS-MS DIA spectra. For DIA spectra, the chromatograms of fragment ions allow comparison of the similarity of their chromatographic peak shapes against each other and against the precursor ion. For LC-IMS-MS DIA spectra, the mobilograms of fragment ions in the additional bottom panel allow comparison of the similarity of their AT peak shapes against each other and against the precursor ion. Furthermore, in this plot the AT offsets of the fragment ions from their precursor, typical for the instrumentation, can be clearly observed. This negative AT shift is a function of the collision energy used and the mass of the fragment ion, and it can be explained by the effect of the accelerating electric field, causing a minimal IMS separation where smaller fragment ions move faster through the collision cell during high-collision energy spectra acquisition, compared to larger fragments and precursor ions[24]. PeakQC automatically generates this kind of figure, which is not available in the handful of existing free IMS-MS tools, and it is very time consuming to generate using the vendor software because it requires several manual steps, one MS run and one fragment ion at a time.

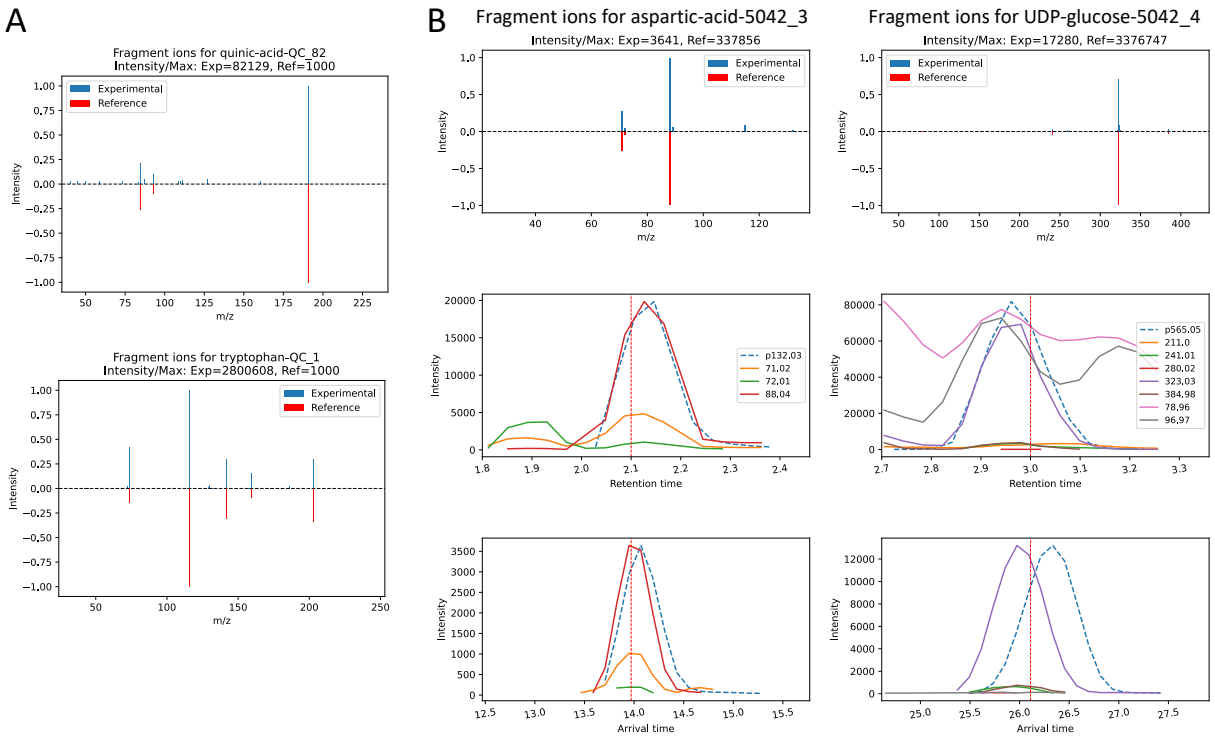

Figure 6. **Example diagnostic figures for MS2 peak signals (fragment ions)**. Mass spectrum with experimental and mirror plot against the reference MS2 (collected in-house from standards), chromatograms, and mobilograms for the fragments of user-specified ions. PeakQC generates

this kind of figures for each user-specified ion and MS run. **(A)** Examples for the metabolomics LC-MS analysis which was performed in DDA mode. Molecules are quinic acid in MS run QC_82 and tryptophan in MS run QC_1. **(B)** Examples for the metabolomics LC-IMS-MS analysis which was performed in DIA mode. Molecules are aspartic acid (*m/z* 132.03, RT 2.1 min, AT 14.0 ms) in MS run 5042_3 and UDP glucose (*m/z* 565.05, RT 3.0 min, AT 26.1 ms) in MS run 5042_4. Blue dashed lines in the chromatograms and mobilograms indicate the trace for the precursor ion with intensity normalized to the maximum fragment ion intensity.

Computing resources required for using PeakQC

The pipeline has low requirements of computing resources and can be run on an average desktop computer with Windows operating system. It takes advantage of parallel processing for some steps according to the number of processing cores available in the computer. Besides the type of computer used, the execution time will vary depending on the number of MS runs, LC duration, and the size of the raw data. For the different datasets in this work, the total execution time for all steps, from raw data conversion to error heatmap generation, ranged from 4 to 45 minutes and used up to 5 GB of random-access memory (on a computer with 48 GB RAM, Intel Core i9–10920X CPU at 3.5 GHz and 64-bit Windows 10 operating system). Specifically, the processing times were 3.97 min for the metabolomics infusion batch of 71 MS runs, 5.97 min for the metabolomics LC-MS batch of 192 MS runs (lower time per file because the original MS1 raw data was stored in centroid mode instead of profile mode), 6.81 min for the metabolomics LC-IMS-MS batch of 8 MS runs, 44.59 min for the lipidomics LC-MS batch of 201 MS runs, and 42.62 min for the proteomics LC-MS batch of 49 MS runs. For large datasets, most of the execution time corresponds to the raw data conversion, but the remaining steps could be re-executed within a few minutes, for example to investigate different tolerances, because the data was already converted.

Conclusions

PeakQC is filling the gap of automated software tools for QC of MS data, especially needed for metabolomics and lipidomics workflows. Our tool performs automated processing and extraction of QC metrics independently of identification tools and omics molecular types, allowing operators to compare multiple MS runs and monitor ions as their data is acquired.

To address limitations of the current version, implementation of a command-line interface and support for more instrument formats are in progress. A command-line functionality will enable two types of important analyses. First, it will enable unattended QC evaluations to perform large-scale comparisons across batches over months and years. And second, it will enable to implement a functionality to monitor a target directory for fully automated re-launch of the PeakQC analysis during data acquisition of a batch of samples as soon as each MS run is acquired, with possibility to automatically send alerting email notifications when defective data is detected. Future improvements could also include additional metrics for MS2 matching (since currently is limited to visualization) and for LC, like full-width-at-half-maximum and peak asymmetry. Moreover, alternative error detection methods will be implemented by leveraging modern AI techniques, which offers an exciting opportunity to reimagine algorithms which operate seamlessly from the MS raw data[7].

We believe that PeakQC represents an advancement in the MS field with emerging tools that work across different omics data types, providing flexibility that enables researchers to perform QC analysis for multiple omics domains within a unified platform, streamlining their workflows and facilitating multiomics data acquisition. PeakQC is enabling collection of high-quality MS data in large-scale studies, avoiding waste of precious samples and resources, and facilitating collection of reproducible and accurate data that can be utilized in downstream analyses to produce high-confidence molecular characterization.

Acknowledgements

Portions of this work were supported by: grant U24CA271012 from the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC); the Agile BioFoundry (http://agilebiofoundry.org) supported by the U.S. Department of Energy (DOE), Energy Efficiency and Renewable Energy, Bioenergy Technologies Office; the Environmental Molecular Sciences Laboratory (EMSL), a DOE Office of Science User Facility sponsored by the Biological and Environmental Research program under Contract No. DE-AC05-76RL01830; and the U.S. DOE, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships (SULI) and Community College Internships (CCI) programs.

Author contributions

Andrea Harrison contributed to software development and manuscript writing; Jennifer E. Kyle, Josie G. Eder, Priscila Lalli, Chaevien S. Clendinen, and Sneha P. Couvillion contributed to conceptualization of software features; Aivett Bilbao conceived and supervised this work, designed and developed the software, and wrote the manuscript. All authors performed or supervised MS experiments, and reviewed and approved the final manuscript.

Notes

The authors declare no competing financial interest.

References

(1) Bittremieux, W.; Willems, H.; Kelchtermans, P.; Martens, L.; Laukens, K.; Valkenborg, D. iMonDB: Mass Spectrometry Quality Control through Instrument Monitoring. *J Proteome Res* **2015**, *14* (5), 2360-2366. DOI: 10.1021/acs.jproteome.5b00127 From NLM Medline.

(2) Schulze, B.; Jeon, Y.; Kaserzon, S.; Heffernan, A. L.; Dewapriya, P.; O'Brien, J.; Ramos, M. J. G.; Gorji, S. G.; Mueller, J. F.; Thomas, K. V.; et al. An assessment of quality assurance/quality control efforts in high resolution mass spectrometry non-target workflows for analysis of environmental samples. *Trac-Trend Anal Chem* **2020**, *133*. DOI: ARTN 116063 10.1016/j.trac.2020.116063.

(3) Bereman, M. S.; Beri, J.; Sharma, V.; Nathe, C.; Eckels, J.; MacLean, B.; MacCoss, M. J. An Automated Pipeline to Monitor System Performance in Liquid Chromatography-Tandem Mass Spectrometry Proteomic Experiments. *J Proteome Res* **2016**, *15* (12), 4763-4769. DOI: 10.1021/acs.jproteome.6b00744 From NLM Medline.

(4) Broeckling, C. D.; Beger, R. D.; Cheng, L. L.; Cumeras, R.; Cuthbertson, D. J.; Dasari, S.; Davis, W. C.; Dunn, W. B.; Evans, A. M.; Fernandez-Ochoa, A.; et al. Current Practices in LC-MS Untargeted Metabolomics: A Scoping Review on the Use of Pooled Quality Control Samples. *Anal Chem* **2023**, *95* (51), 18645-18654. DOI: 10.1021/acs.analchem.3c02924 From NLM Medline.

(5) Dunn, W. B.; Kuligowski, J.; Lewis, M.; Mosley, J. D.; Schock, T.; Ulmer Holland, C.; Zanetti, K. A.; Vuckovic, D.; Metabolomics Quality, A.; Quality Control, C. Metabolomics 2022 workshop report: state of QA/QC best practices in LC-MS-based untargeted metabolomics, informed through mQACC community engagement initiatives. *Metabolomics* **2023**, *19* (11), 93. DOI: 10.1007/s11306-023-02060-4 From NLM Medline.

(6) Gonzalez-Dominguez, A.; Estanyol-Torres, N.; Brunius, C.; Landberg, R.; Gonzalez-Dominguez, R. QComics: Recommendations and Guidelines for Robust, Easily Implementable and Reportable Quality Control of Metabolomics Data. *Anal Chem* **2024**, *96* (3), 1064-1072. DOI: 10.1021/acs.analchem.3c03660 From NLM Medline.

(7) Ross, D. H.; Bhotika, H.; Zheng, X.; Smith, R. D.; Burnum-Johnson, K. E.; Bilbao, A. Computational tools and algorithms for ion mobility spectrometry-mass spectrometry. *Proteomics* **2024**, e2200436. DOI: 10.1002/pmic.202200436 From NLM Publisher.

(8) Bilbao, A.; Varesio, E.; Luban, J.; Strambio-De-Castillia, C.; Hopfgartner, G.; Muller, M.; Lisacek, F. Processing strategies and software solutions for data-independent acquisition in mass spectrometry. *Proteomics* **2015**, *15* (5-6), 964-980. DOI: 10.1002/pmic.201400323 From NLM Medline.

(9) Sandhu, W.; Gray, I. J.; Lin, S.; Elias, J. E.; DeFelice, B. C. Rapid QC-MS – Interactive Dashboard for Synchronous Mass Spectrometry Data Acquisition Quality Control. *bioRxiv* **2024**, 2024.2001.2030.578059. DOI: 10.1101/2024.01.30.578059.

(10) Wallmann, G.; Leduc, A.; Slavov, N. Data-Driven Optimization of DIA Mass Spectrometry by DO-MS. *J Proteome Res* **2023**, *22* (10), 3149-3158. DOI: 10.1021/acs.jproteome.3c00177 From NLM Medline.

(11) An, S.; Wang, R.; Lu, M.; Zhang, C.; Liu, H.; Wang, J.; Xie, C.; Yu, C. MetaPro: a web-based metabolomics application for LC-MS data batch inspection and library curation. *Metabolomics* **2023**, *19* (6), 57. DOI: 10.1007/s11306-023-02018-6 From NLM Medline.

(12) Kopczynski, D.; Hoffmann, N.; Troppmair, N.; Coman, C.; Ekroos, K.; Kreutz, M. R.; Liebisch, G.; Schwudke, D.; Ahrends, R. LipidSpace: Simple Exploration, Reanalysis, and Quality Control of Large-Scale Lipidomics Studies. *Anal Chem* **2023**, *95* (41), 15236-15244. DOI: 10.1021/acs.analchem.3c02449 From NLM Medline.
(13) Morgenstern, D.; Barzilay, R.; Levin, Y. RawBeans: A Simple, Vendor-Independent, Raw-Data Quality-Control Tool. *J Proteome Res* **2021**, *20* (4), 2098-2104. DOI: 10.1021/acs.jproteome.0c00956 From NLM Medline.
(14) Stanfill, B. A.; Nakayasu, E. S.; Bramer, L. M.; Thompson, A. M.; Ansong, C. K.; Clauss, T. R.; Gritsenko, M. A.; Monroe, M. E.; Moore, R. J.; Orton, D. J.; et al. Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data. *Mol Cell Proteomics* **2018**, *17* (9), 1824-1836. DOI: 10.1074/mcp.RA118.000648 From NLM Medline.
(15) Amidan, B. G.; Orton, D. J.; Lamarche, B. L.; Monroe, M. E.; Moore, R. J.; Venzin, A. M.; Smith, R. D.; Sego, L. H.; Tardiff, M. F.; Payne, S. H. Signatures for mass spectrometry data quality. *J Proteome Res* **2014**, *13* (4), 2215-2222. DOI: 10.1021/pr401143e From NLM Medline.
(16) Taylor, R. M.; Dance, J.; Taylor, R. J.; Prince, J. T. Metriculator: quality assessment for mass spectrometry-based proteomics. *Bioinformatics* **2013**, *29* (22), 2948-2949. DOI: 10.1093/bioinformatics/btt510 From NLM Medline.
(17) Pichler, P.; Mazanek, M.; Dusberger, F.; Weilnbock, L.; Huber, C. G.; Stingl, C.; Luider, T. M.; Straube, W. L.; Kocher, T.; Mechtler, K. SIMPATIQCO: a server-based software suite which facilitates monitoring the time course of LC-MS performance metrics on Orbitrap instruments. *J Proteome Res* **2012**, *11* (11), 5540-5547. DOI: 10.1021/pr300163u From NLM Medline.
(18) Bilbao, A.; Ross, D. H.; Lee, J. Y.; Donor, M. T.; Williams, S. M.; Zhu, Y.; Ibrahim, Y. M.; Smith, R. D.; Zheng, X. MZA: A Data Conversion Tool to Facilitate Software Development and Artificial Intelligence Research in Multidimensional Mass Spectrometry. *J Proteome Res* **2023**, *22* (2), 508-513. DOI: 10.1021/acs.jproteome.2c00313 From NLM Medline.
(19) Kyle, J. E.; Crowell, K. L.; Casey, C. P.; Fujimoto, G. M.; Kim, S.; Dautel, S. E.; Smith, R. D.; Payne, S. H.; Metz, T. O. LIQUID: an-open source software for identifying lipids in LC-MS/MS-based lipidomics data. *Bioinformatics* **2017**, *33* (11), 1744-1746. DOI: 10.1093/bioinformatics/btx046 From NLM Medline.
(20) Birer-Williams, C. M. C.; Chu, R. K.; Anderton, C. R.; Wright, E. S. SubTap, a Versatile 3D Printed Platform for Eavesdropping on Extracellular Interactions. *mSystems* **2021**, *6* (4), e0090221. DOI: 10.1128/mSystems.00902-21 From NLM PubMed-not-MEDLINE.
(21) Bilbao, A. *PeakDecoder enables machine learning-based metabolite annotation and accurate profiling in multidimensional mass spectrometry measurements*. 2023. (accessed.
(22) Bilbao, A.; Gibbons, B. C.; Stow, S. M.; Kyle, J. E.; Bloodsworth, K. J.; Payne, S. H.; Smith, R. D.; Ibrahim, Y. M.; Baker, E. S.; Fjeldsted, J. C. A Preprocessing Tool for Enhanced Ion Mobility-Mass Spectrometry-Based Omics Workflows. *J Proteome Res* **2022**, *21* (3), 798-807. DOI: 10.1021/acs.jproteome.1c00425 From NLM Medline.
(23) Liu, F. T.; Ting, K. M.; Zhou, Z. H. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, 15-19 Dec. 2008, 2008; pp 413-422. DOI: 10.1109/ICDM.2008.17.
(24) Mairinger, T.; Kurulugama, R.; Causon, T. J.; Stafford, G.; Fjeldsted, J.; Hann, S. Rapid screening methods for yeast sub-metabolome analysis with a high-resolution ion mobility

quadrupole time-of-flight mass spectrometer. *Rapid Commun Mass Spectrom* **2019**, *33 Suppl 2* (Suppl Suppl 2), 66-74. DOI: 10.1002/rcm.8420 From NLM Medline.