How much storage precision can be lost: Guidance for near-lossless compression of untargeted metabolomics mass spectrometry data.

Junjie Tong [1], Miaoshan Lu [2], Bichen Peng [1], Shaowei An [3], Jinyin Wang [2] and Changbin Yu [1,*]

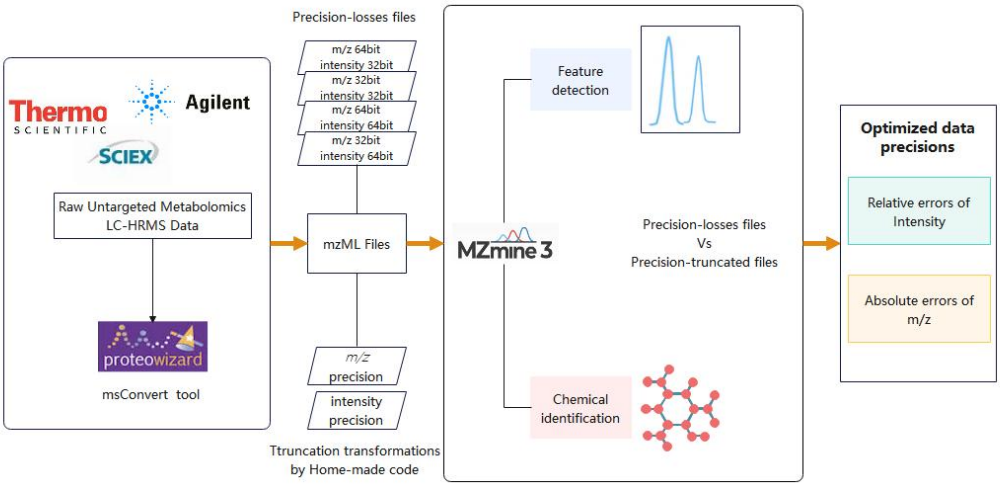
1 Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan 250000, China; tongjunjie@sdfmu.edu.cn

2 Zhejiang University; lumiaoshan@westlake.edu.cn

3 Fudan University; anshaowei@westlake.edu.cn

* Correspondence: yu_lab@sdfmu.edu.cn;

Graphic Abstract



Abstract:

Backgroud: The size of high-resolution mass spectrometry (HRMS) data has been increasing significantly. Several lossy compressors have been developed for higher compression rate. Currently, a comprehensive evaluation of what and how MS data (*m/z* and intensities) with precision losses would affect data processing is absent. Assessing the impact of different degrees of precision losses on the data processing results should clarify the variation rates under different accuracy losses and explore the reasons for them.

Result: Sixteen vendor files were converted to mzML files with a different combination of data precision (32- or 64-bit) for *m/z* and intensities via msConvert. A suitable precision combination of mzML files were afterwards converted to precision-lossy files with absolute *m/z* or relative intensities mistakes by truncation transformations. We set an error threshold at 1% to evaluate files results of feature and compound detection obtained from MZmine3. The variation was <0.13% for both features and compounds when *m/z* and intensities with different combinations of storage precision. Five maximum absolute errors of *m/z* (10^{-5}, $2x10^{-5}$, $5x10^{-5}$, 10^{-4}, 10^{-3}) and five maximum relative errors of intensities ($2x10^{-4}$, $2x10^{-3}$, $8x10^{-3}$, $2x10^{-2}$,

$2x10^{-1}$) were examined. We identified that the error of 10^{-4} for *m/z* had a feature detection error of 0.57% and compound detection error of 1.1%. For intensities, the error group of $2x10^{-2}$ had an error of 4.65% for features and 0.98% for compounds to precision-lossless files. Taken together, we consider that a maximum absolute error of 10^{-4} for *m/z* and a maximum relative error of $2x10^{-2}$ for intensity can meet the error threshold of 1% and be recommended errors for lossy compression.

Conclusion: We examined that mzML files with both *m/z* and intensity encoded in 32-bit precision appear to be a preferred combination, which has smaller file size and minor variation. Further, we checked that how varying levels of precision affect the MS data processing and provided a reasonable scene-accuracy proposal (10^{-4} for *m/z* and $2x10^{-2}$ for intensities). This guidance aimed to help researchers in improving lossy compression algorithms and minimizing the negative effects of precision losses on downstream data processing.

Keywords: HRMS data, Precision Control, Metabolomics, Data Compression.

1. Introduction

Untargeted metabolomics, as a study of measuring the metabolome within an organism, has been extensively applied for disclosing novel biomarkers clinically [1]. In this field of study, liquid chromatography-mass spectrometry (LC-MS) technology is considered the prime choice for small molecule identification and quantification. To exploit the information from a raw MS output file requires a workflow of data analysis, including data processing, data annotation, statistical analysis, functional analysis, etc., [2]. Among them, data processing is a critical step and the cornerstone of untargeted metabolomics data analysis, which comprises noise signal filtration, extracted ion chromatogram (EIC), peak detection, peak deconvolution, and peak alignment [2] [3]. Several bioinformatics tools and platforms have been introduced to address the needs of processing untargeted metabolomics data [4], some of which are open-source software, like MZmine3 [5, 6], MS-Dial [7], XCMS [8], MetaboAnalyst 5.0 [1], and other commercial tools, such as Compound Discoverer (Thermo Scientific), MarkerView (SCIEX), Mass Profiler Professional (Agilent), MarkerLynx (Waters).

Raw files from the mass spectrometer will be automatically stored in a corresponding vendor proprietary format, for example, the .Raw extension for Thermo Scientific and .wiff extension for SCIEX instruments [9]. Due to the small file size, these vendor files are suitable for long-term storage but limited in data exchange and distribution across the MS community as its nature of lacking cross-platform compatibility and software flexibility. As a result, a community standard format, "mzML", has been developed and rapidly become the most commonly open format for MS data [10]. One of the disadvantages of mzML is the much bigger file size than vendor format, resulting in a higher cost of storage. Thus, a number of compression algorithms for MS data have been proposed recently. In general, these compression strategies could be classified into lossless, such as MassComp [11], and lossy (or "near-lossless"), like MSNumpress [12], Aird-ZDPD [13], or both, like mspack [14].

These lossy algorithms achieved higher compression rates at the expense of certain storage precision of raw data. The following question is how much precision losses of raw data (i.e., *m/z* and intensities) will affect the downstream data processing to some acceptable or unacceptable extent. In MSNumpress, Teleman et al. reported a maximum relative error of $2x10^{-9}$ (0.002ppm) for *m/z* and $2x10^{-4}$ for intensities generated the identical peptide list (FDR < 1%) as the raw files. For mspack, Hanau et al. showed that a compressed file with a maximum absolute error of 10^{-4} (0.2ppm at 500 *m/z*) for *m/z* and a relative error of 10^{-2} for intensities

would be well below the mass accuracy of mass analyzers and the difference between repetitions of an experiment, respectively. In a recent study, Wang et al. developed Stack-ZDPD from ZDPD and observed that the *m*/*z* data stored at 5 decimal places (0.1ppm at 100 *m*/*z*) generate the almost same peak lists detected via MZmine3 [15]. The precision losses of *m*/*z* values and intensities without negative impacts on data analysis vary in different compressors, which may be caused by different output file formats (MSNumpress for mzML, mspack for both mzML and mzXML, Stack-ZDPD for Aird format) and different data processing software were applied. Till now, there is no study has systematically evaluated the impacts of data precision losses of *m*/*z* and intensities data in mzML files on subsequent data processing step.

In this study, we first converted metabolomics MS vendor files into precision-lossless mzML files via command line tools of msConvert [16], where *m*/*z* values and intensities were encoded in four combinations of storage precision (32- or 64-bit). Then, we tested the differences of these combinations based on the critical results of feature detection and compound identification during data processing, with the former describing the processing results of the original signal while the latter is more dependent on the scope of the search library. We used a two-dimensional window of RT shift <= 0.025min and *m*/*z* error <= 5ppm to determine whether two features are identical. Furthermore, we generated precision-lossy mzML files with a defined maximum absolute error of *m*/*z* or a relative error of intensities by the truncation transformations mentioned in mspack. These truncated mzML files were also tested in terms of feature and compound detection. Eventually, we compared precision-lossy files to precision-lossless files in the total number of features, consensus features ratio, intensity fold change (FC) of consensus features, and annotated consensus compounds.

Considering the mass accuracy of current instruments and reproducibility of an experiment, we set the threshold value of the error to 1%. After comparing the above differences with the error threshold, we proposed a guidance for a suitable storage precision combination in precision-lossless mzML files, and for the storage precision losses of *m*/*z* and intensity value in precision-lossy compression.

2. Materials and Methods

2.1 Vendor Files Preparation

In order to complete a comprehensive evaluation, we selected sixteen LC-MS vendor files (labeled from Met_01 to Met_16) generated by different types of instruments, from various organisms, and acquired by different acquisition methods (Table 1). All vendor files can be searched on the open-access database repository of Metabolights (https://www.ebi.ac.uk/metabolights/) by their original file name in Table 1.

To determine the suitability of files, we required them must meet the inclusion criteria. (1), It must be the raw vendor file rather than the derived file format (e.g., mzML, mzXML), because the storage precision of mzML files has been fixed. (2), These files were generated from untargeted metabolomics or lipidomics studies as the workflow of MZmine3 is designed for untargeted metabolomics data processing. (3), The MS data from high-resolution mass spectrometers is required as the mass detection algorithm is suitable for HRMS data. Finally, sixteen vendor files were included and mainly from two types of analyzers, time-of-flight (TOF) and orbitrap. In terms of acquisition modes, thirteen of them were acquired by data-dependent acquisition (DDA), one by data-independent acquisition (DIA), and the remaining two by all ion fragmentation (AIF). For the origins of these files, nine of them are human samples (plasma, urine, Hep-G2 cell), three are from yeast and four from the

plant. To better evaluate the effects of *m*/*z* and intensity errors broadly, we considered a wide range of file sizes, ranging from 38.5MB to 1.21GB.

Table 1. Raw MS vendor files used for evaluating the effects of storage precision on data processing steps. The information of file name, vendor instruments, data acquisition type and organisms in each file is included. Met_01 and 02 are from [17]. Met_03-06 are from [18]. Met_07 is from [19]. Met_08-09 are from [20]. Met_10 is from [21]. Met_11 is from [22]. Met_12 is from [23]. Met_13-14 are from [24]. Met_15-16 are from [25].

File No.	Original File Name	Vendors Name	Instrument	Data Acquisition Type	Organism (part)
Met_01	Plasma3PlasmaDDA20-50.wiff	SCIEX	Triple TOF6600	DDA	Homo sapiens (Blood Plasma)
Met_02	PestMix-1-3-5-7Step1Plasma3SWATH20-50.wiff	SCIEX	Triple TOF6600	DIA	Homo sapiens (Blood Plasma)
Met_03	SA1.raw	Thermo	Q Exactive HF	DDA	Piper nigrum (Seed)
Met_04	SB1.raw	Thermo	Q Exactive HF	DDA	Piper nigrum (Seed)
Met_05	SampleA_1.wiff	SCIEX	Triple TOF6600	DDA	Piper nigrum (Seed)
Met_06	SampleB_1.wiff	SCIEX	Triple TOF6600	DDA	Piper nigrum (Seed)
Met_07	arunJ_20210318_o24433_pos_AIF_SOP4_empvec_1.raw	Thermo	Q Exactive	AIF	Saccharomyces cerevisiae (Whole Organism)
Met_08	04102020_Covid_TP_Mild_9030_1_P325_P_2.raw	Thermo	Q Exactive	DDA	Homo sapiens (Blood Plasma)
Met_09	23092020_Covid_TP_Severe_9030_2_P411_P_2.raw	Thermo	Q Exactive	DDA	Homo sapiens (Blood Plasma)
Met_10	TFHN1.raw	Thermo	Q Exactive HF	DDA	Homo sapiens (Blood Plasma)
Met_11	180327_S1_LC02MS02_04_A12_ProlineBetaine.d	Agilent	6550 Q-TOF	AIF	Homo sapiens (Urine)
Met_12	016_HEP_SOL1_LYS_neg_Spiking.raw	Thermo	Q Exactive HF	DDA	Homo sapiens (Hep-G2 cell)
Met_13	CE04.raw	Thermo	Exactive Orbitrap	DDA	Homo sapiens (Urine)
Met_14	CE05.raw	Thermo	Exactive Orbitrap	DDA	Homo sapiens (Urine)
Met_15	Positive_000324.raw	Thermo	Q Exactive Plus	DDA	Saccharomyces cerevisiae (Whole Organism)
Met_16	Negative_000333.raw	Thermo	Q Exactive Plus	DDA	Saccharomyces cerevisiae (Whole Organism)

2.2 Precision-lossless Files Encoded in 32-bit and 64-bit precision

Sixteen vendor files were converted to the open format, mzML, by the command line tool of msConvert (ProteoWizard) [16]. Each vendor file generated four mzML files with *m/z* and intensities encoded in different combinations of storage precision: (1), both *m/z* and intensity values were stored at single (32-bit) precision, with parameters "--mz32 --inten32" in msConvert. (2), both *m/z* and intensity data with double (64-bit) precision, parameters as "--mz64 --inten64". (3), *m/z* and intensities encoded in 32-bit and 64-bit precision, respectively, parameters as "--mz32 --inten64". (4), *m/z* at 64-bit precision and intensities at 32-bit precision ("--mz64 --inten32").

2.3 Truncation Transformations for m/z and Intensity Values

The truncation transformations idea is one of the lossy compression methods used in mspack compressors [14]. According to the IEEE 754 standard format of a floating-point number (Figure1), the single-precision floating numbers contained one sign bit, eight exponent bits, and 23 fraction bits, for example, a binary format of 32-bit floating point number could be expressed by $(-1)^{sign}2^{exponent}$x1.*fraction*. Thus, a single-precision floating number has 24 bits (one more than fraction bits) of precision and a double-precision floating number has 53 bits of precision. The idea of truncation transformation is to set the last significant bits in the fraction part to zero, which creates a certain degree of precision losses on the import value. In the MS data, *m/z* and intensity values could be transformed to data with defined errors by this idea.

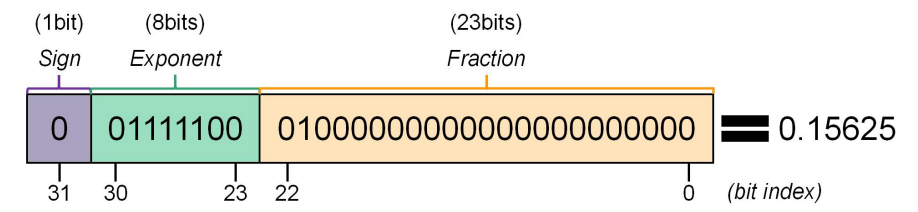


Figure 1. This figure is adopted from [13], which shows the binary format of value 0.15625 in IEEE 754 standard with single precision. The value can be computed from the binary representation as $(-1)^{b_{31}}\left(1+\sum_{i=1}^{23} b_{23-i}2^{-i}\right)2^{(e-127)}$, with $e=\sum_{i=0}^{7} b_{23+i}2^{i}$.

Next, we transformed these precision-lossless files from Chapter 2.2 and created lossy files with a certain degree of maximum absolute error of *m/z* or relative error of intensity data. In most representations of MS data, *m/z* is store at absolute value, but intensities are shown as relative intensity, the ratio of intensities of resolved peaks to the intensity of base peak (the most intense peak). Thus, we chose the absolute error for *m/z* and relative error for intensities as the same in mspack [15]. The precision-lossy files were used for the downstream data processing workflow. This homemade truncation transform code, named PrecisionControlTask, was written in Java language and was locally integrated into the IO module (import_rawdata_mzml) of a forked Java-based MZmine3 source code (https://github.com/TongJJCSi/mzmine3).

2.3.1 The truncation transformations for *m/z* values (absolute errors)

The input *m/z* floating-point values (32-bits or 64-bits) were converted to their binary format under IEEE754 standard. Based on the defined absolute error, we calculated the fraction bits needed to be truncated (details of bits calculation can be found here [14]) and set them to zero. Subsequently, each precision-lossless mzML file was transformed to five precision-lossy mzML files with different degrees of absolute errors in *m/z*, resulting in

maximum absolute errors of 10^{-5}, $2x10^{-5}$, $5x10^{-5}$, 10^{-4} and 10^{-3}. The sixteen mzML files with the same absolute errors were regarded as an error group. These error groups were selected with reference to previous studies, which showed that MSNumpress ($2x10^{-9}$ or 0.002ppm) and mspack (10^{-4} or 0.2ppm at 500 *m/z*) defined the maximum relative and absolute errors for *m/z* without negative effects on the following data processing. Therefore, we first chose an intermediate absolute error of 10^{-5} (0.02ppm at 500 *m/z*) that was examined no difference with precision-lossless files (error threshold < 1%), and then extended error range to 10^{-3}, where significant difference between precision-lossless and -lossy files were occurred.

2.3.2 The truncation transformations for intensity values (relative errors)

Similar to the truncation transformation used for *m/z* values. In order to control the maximum relative error of intensities, the way of calculating the number of truncated bits was referenced to the idea in mspack [14]. Then, we set five error groups with a range of maximum relative errors, $2x10^{-4}$, $2x10^{-3}$, $8x10^{-3}$, $2x10^{-2}$ and $2x10^{-1}$. All files with the same relative errors were clustered into an error group. The reason for setting these groups is similar to *m/z* Based on the maximum relative errors in MSNumpress ($2x10^{-4}$) and mspack (10^{-2}), we checked the range of relative errors from $2x10^{-4}$ to $2x10^{-1}$, describing the trend of variation in feature and compound identification is getting bigger as the expansion of maximum relative errors of intensity data.

In general, each precision-lossless file was transformed into ten lossy files (five with only *m/z* errors and five with only intensity errors)

2.4 LC-MS Data Processing Workflow via MZmine3

MZmine 3 (v3.2.8) was used to process both precision-lossless (files from Chapter 2.2) and precision-lossy (files from Chapter 2.3) mzML files. All parameters used in MZmine3 can be seen in Table S1(See Supplementary). All mzML files were imported into MZmine3 in profile mode (except Met-11 in centroided mode), so the first step was the mass detection module for data-centroiding. In this module, the algorithm of exact mass was used to centroid the profile data. In order to obtain more peaks, we set the noise level at 1. The noise level set a noise threshold and excluded background noise under the defined level, which is dependent on the type of mass spectrometers. Thus, the recommended way of setting noise level is reviewing each file, but it may still miss some signals. Therefore, we set the value at 1 to include more features and set another noise-filter downstream.

Afterwards, the EIC building was processed by the Automated Data Analysis Pipeline (ADAP) chromatogram builder, where the scan-to-scan accuracy (*m/z*) parameter was set to 10ppm for orbitrap LC-MS data and 15ppm for time-of-flight (TOF) generated files. Then, peaks in these EICs were detected by ADAP feature resolver module with default parameters, where Signal-to-Noise was estimated in an intensity window-based method. The feature alignment was performed in a default setting as well.

After all, each feature list was searched against an online database via the integrated annotation module in MZmine3. The determination of the database was based on the organisms ' type of mzML files, such as the human samples were against Human Metabolome Database (HMDB) (Met_01, 02, 08, 09, 10, 11, 12, 13, 14), the yeast sample against Yeast Metabolome Database (YMDB) (Met_15), and the remaining samples against Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolites database (Met_03, 04, 05, 06, 07, 16). The feature lists annotated by the metabolites were exported as .csv format and stored locally for further data statistics.

2.5 Data Statistics of Precision-lossless Files

The feature list and annotated compound list of each precision-lossless file (from Chapter 2.2) were exported via MZmine3. For feature lists, we defined a feature via its *m/z* and retention time (RT) value. Then, we examined the overlapping features (consensus features) of four files and presented them as Venn diagrams, which are generated from the same vendor file but with different combinations of storage precision. Next, we calculated the proportion of unique features for each file to demonstrate the difference between these combinations. To avoid mistakenly assigning "real" consensus features with a minor margin of error as a unique feature, we applied a window of *m/z* <= 5ppm and RT shift <= 0.025min to filter consensus features. Based on the different types of instruments, we further filtered out noises with a recommendation from software documentation for every file, i.e., 1.0E2 for TOF analyzers and 1.0E4 for Orbitrap. After the filtration, we recalculated the filtered fraction of unique features for each file.

Downstream of feature detection, we defined a compound through its compound name, molecular formular, and neutral mass information from compound lists. In each file, we counted the total number of annotated compounds and screened out duplicate compounds. Afterward, we calculated the unique compound rate (UCR) in all files.

After all, we evaluated the differences in feature and compound detection between storage precision combinations via the threshold value of error (1%).

2.6 Data Statistics of Truncated Files

All *m/z* and intensity precision-truncated files (from Chapter 2.3) were processed via MZmine3 and generated feature and compound lists. The feature and compound detection results of truncated files were parallelly compared to the results of precision-lossless files (from Chapter 2.5). The overlapping features or compounds between precision-lossy and -lossless files were regarded as consensus features or compounds while these only detected in precision-lossy files are called unique features or compounds.

In feature lists, the number of total features and consensus features were calculated. Then, we checked the filtration results of consensus features under a window of *m/z* <= 5ppm and RT shift <= 0.025min. After filtration, we calculated the ratio of unique features to total features of precision-lossless files, which was compared to the acceptable margin of error (1%). Furthermore, we measured the distribution of fold-change (FC) of intensity in consensus features. In compound lists, the total number and percentage of consensus compounds across all files were calculated to evaluate the impacts of different degrees of precision losses on compound detection.

3. Results and Discussion

3.1 Single- Vs Double-Precision Storage for m/z and Intensities in mzML Files

Currently, mzML is one of the most popular open formats of MS files and is widely used in open-source untargeted metabolomics software. Converting raw vendor files to mzML files using msConvert with default settings would encode the *m/z* values in 64-bit precision (i.e., double-precision) and intensity values in 32-bit precision (i.e., single-precision) for the output files [12]. However, in some data processing software, both *m/z* and intensity data are calculated in double precision. Therefore, we aim to test and evaluate the impact of the degree of variation of mzML files at four different combinations (mz32- or 64-bits, intensity32- or -64bits) of precision on the subsequent data processing. Here, these files were imported into MZmine3 with defined parameters, where feature list and annotated compound list were output as results. The lists of features and compounds represented the most important steps during data analysis, which can clearly elucidate how the different combinations of precision have a variation on the final compound identification results.

3.1.1 Comparison of Feature Detection Results

All 64 mzML files (16 vendor files times 4 types of data precision combinations) were processed with the parameters mentioned in Table S1, exporting 64 feature lists. A feature is defined as a two-dimensional signal (i.e., *m*/*z* and retention time) [26]. Thus, we recognized a non-zero feature with identical *m*/*z* and retention time (RT) overlapping across multiple mzML files to be a consensus feature. From the Venn diagram of detected features, we observed the number of consensuses features from 9 (Met_10) to over 46,000 (Met_03), demonstrating the complexity of the selected 16 mzML files (Figure 2). When *m*/*z* values are encoded in a fixed precision, files have the identical feature list regardless of storage precision of corresponding intensity values. However, the storage precision of intensity data at a fixed point, *m*/*z* values would generate a feature list with differences between 32-bit and 64-bit precision. These features, which were detected only in one specific *m*/*z* data precision, were called unique features. In Met_01 file, for example, the mz32intensity32 (403 features detected) shared all 403 features with the mz32intensity64 file, but only 386 features are same with the mz64intensity32. Thus, the Met_01 replicates had 386 consensus features and 17 (*m*/*z* 32-bits files) and 15 (*m*/*z* 64-bits files) unique features.

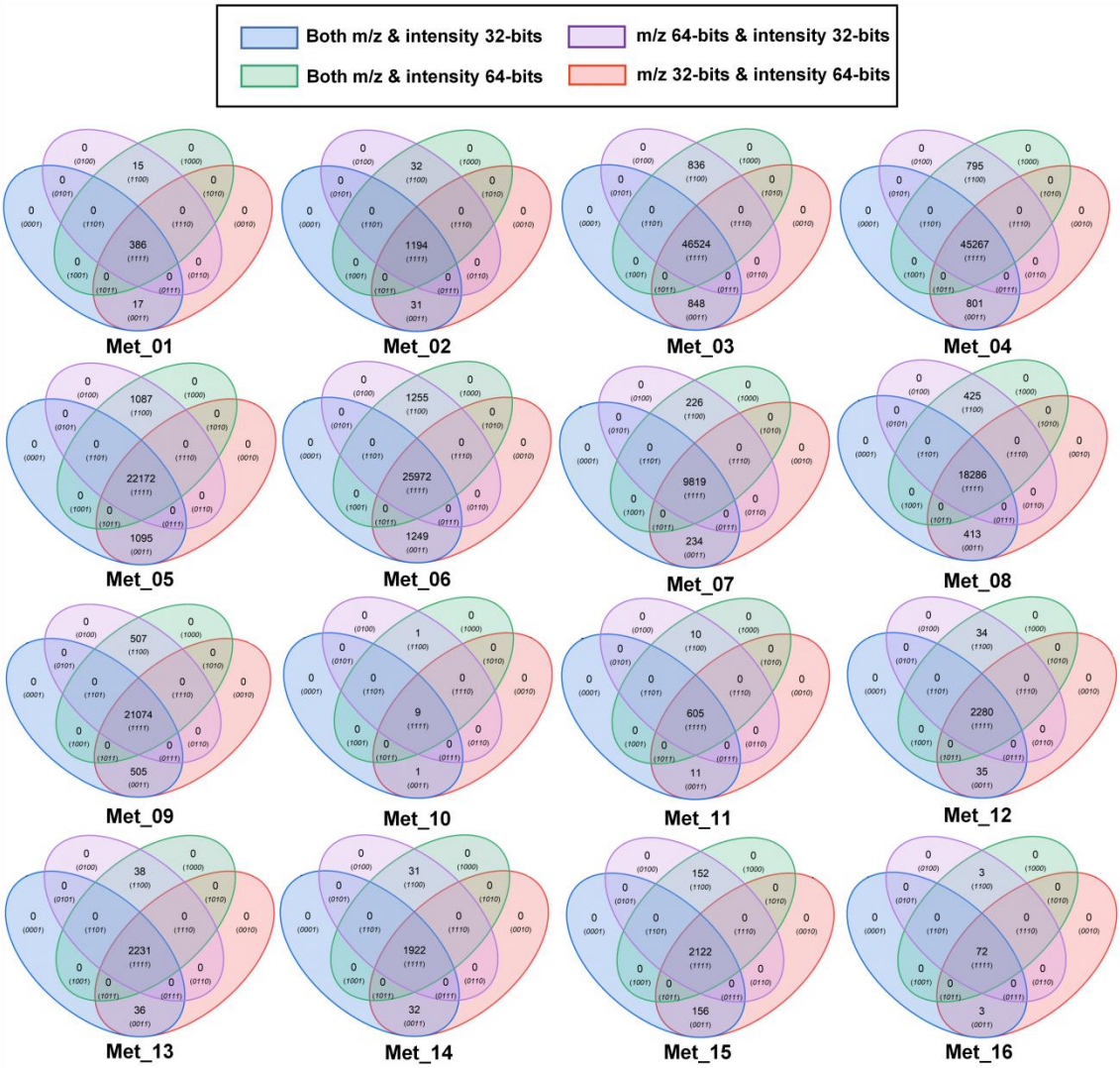


Figure 2. Comparison of detected features of different storage precision combinations across 16 files. both *m*/*z* and intensity for 32-bit precision (Blue), both *m*/*z* and intensity for 64-bit precision (Green), *m*/*z* for 64-bit and intensity for 32-bit precision (Purple), *m*/*z* for 32-bit and intensity for 64-bit precision (Red).

We calculated the unique feature rate (UFR), the percentage of unique features to total features, to evaluate the degree of difference between files with *m*/*z* at 32- or 64-bits precision. Under the definition of consensus features (*m*/*z* and RT values are identical), the average UFR between files with fixed precision of intensity and alternative precision of *m*/*z* values (32- or 64-bits) is about 3.3% (see unfilter lines in Figure3a). This unfiltered UFR showed a noticeable degree of difference (>1% of threshold value of error) in feature lists of files.

However, we noticed that a part of unique features with a minor margin of difference to consensus features and some background noises have not been removed. A feature occurs across repetitions of an experiment may bring a RT drift (few seconds to tens of seconds) problem, or *m*/*z* errors due to mass analyzers (e.g., 1-5ppm for orbitrap) [27, 28]. Some studies suggested that RT shift and *m*/*z* errors could be corrected via a certain two-dimensional window of *m*/*z* and RT, such as the window of RT shift < 1.5s (0.025min) and *m*/*z* error < 0.01 (20ppm at 500 *m*/*z*) [29], or a window of RT shift < 0.3min and *m*/*z* error < 10ppm [18]. We recognized that the RT shift, caused by the changes of storage precision in the same file, should be minor, which has also a limited impact on the feature detection since RT only has an affect when determining a peak via a feature determination algorithm. Thus, we set the RT shift less than 1.5s (0.025min). Then, we examined the distribution of unique features, detected from all *m*/*z* 32- and 64-bit precision files, over different relative error (ppm) ranges under the RT shift of 0.025min (Figure 3b). We identified that over 80% of unique features in all files could be reassigned to consensus features with a window of *m*/*z* error <= 1 ppm and continuing to expand this window to 5ppm the proportion can increase to 90%. However, we further expand this window to 10ppm has a limited help in reassigning more consensus features. Therefore, we used a window of RT <= 0.025min and *m*/*z* error <= 5 ppm to screen all unique features. For the noise filtration, MZmine3 documentation suggested a noise threshold of 1.0E2 for TOF analyzers and 1.0E4 for Orbitrap instruments. Thus, the remaining unique features were subsequently filtered out these features with intensity values lower than 1.0E2 (TOF analyzers, SCIEX & Agilent instruments) and lower than 1.0E4 (Orbitrap analyzers, Thermo instruments) (Figure 3c, 3d). Consequently, the filtered UFR were adjusted to a negligible level across all files, with an average rate of 0.13% (see filtered lines in Figure 3a).

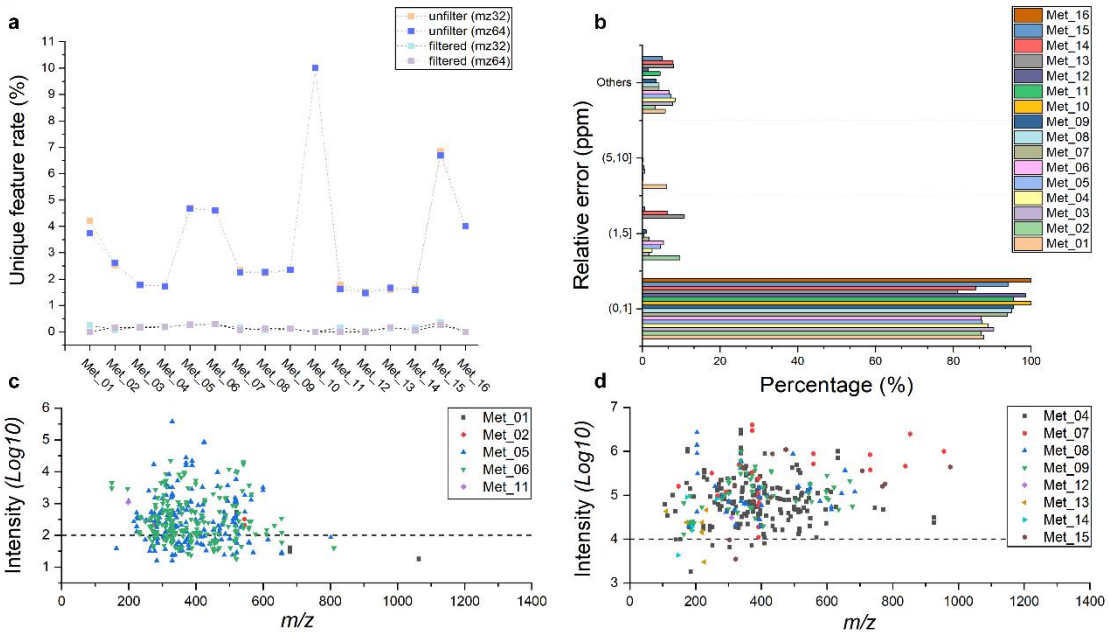


Figure 3. The filtration process of unique features in both 32-bit and 64-bit precision mzML files. (**a**) The unique feature rate in all files before and after filtration; (**b**) The distribution of relative errors of unique

features in *m/z* 32-bit and 64-bit replicates of all 16 files within a RT shift <= 0.025min. The group of "others" comprises features with RT shift > 0.025min or *m/z* error > 10 ppm; (c), (d) The distribution of *m/z* and intensity values in all unique features (c for TOF analyzers, d for Orbitraps) filtered by a window of RT <= 0.025min and *m/z* error <= 5ppm.

From the point of view of feature lists, we recognized that the filtered UFR (0.13%) is well below the acceptable error threshold (1%) that we set. However, the discrepancy needed to be further evaluated on the compound annotation list to determine if the difference in feature lists carries over to the list of identified compounds. As files with fixed storage precision of *m/z* and 32-bit or 64-bit precision of intensities generate the same feature lists, we then evaluated differences between files with *m/z* values of 32-bit and 64-bit precision in the following compound identification.

3.1.2 Comparison of Online Compound Database Search Results

Feature lists of all files were used to match against online compound databases based on the sample origins via the online compound database search module of MZmine3. The total number of detected compounds ranges from 3 to 9765 among sixteen files (Table S2). A compound is defined by its compound name, molecular formula, and neutral mass. After screening out these repeated compounds, the number of compounds saw a decrease. Next, we calculated the UCR for all files, except for Met_10, Met_15, and Met_16 files with a compound number less than 50, resulting in an average UCR of 0.13%.

Overall, we considered that the variation on feature (0.13%) and compound (0.13%) detection across different combinations of *m/z* and intensity storage precision is well below the error threshold (1%). Thus, the files with both *m/z* and intensity values encoded in 32-bit precision would be a preferred choice when researchers require smaller file sizes and faster data analysis. In the following study, we selected these files (mz32intensity32) for truncation transformation to examine a suitable precision loss for *m/z* and intensity value.

3.2 Truncation Transform of Intensity Values

Previous studies showed that files compressed via certain compression algorithms with a maximum relative error of $2x10^{-4}$ (MSNumpress) or 10^{-2} (mspack) for intensity values would not affect results of the following data processing steps [12, 14]. Thus, we aimed to test the changes of feature and compound lists as the relative errors of intensities increase. Here, we set five groups ($2x10^{-4}$, $2x10^{-3}$, $8x10^{-3}$, $2x10^{-2}$, $2x10^{-1}$) by truncation transformation to identify the point of relative error with or without changes for data processing.

3.2.1 Comparison of Feature Lists among Intensity Groups with Different Relative Errors

To assess the differences between groups, we set the precision-lossless files as the "Blank" group, Features shared between precision-lossless and lossy files were considered as " Consensus features " , features detected only in precision-lossy files were regarded as "Unique features".

The number of total features had some small fluctuations between blank and error groups ($2x10^{-4}$, $2x10^{-3}$, $8x10^{-3}$, and $2x10^{-2}$), while a significant decrease occurred in the group of all files with a relative error of $2x10^{-1}$ ($p < 0.05$) (Table S3). For the number of consensus features, a noticeable decrease started in the group of $8x10^{-3}$, where thousands of unique features were identified in some files, culminating in an average of less than 20 percent of consensus features in group $2x10^{-1}$ (Figure 4).

As we mentioned above, an appropriate two-dimensional window of RT and *m/z* shift could reassign "fake" unique features back to consensus features. To explore the degree of differences in feature detection between these errors groups correctly, we calculated the ratio of consensus features to total features in the blank file (except for Met_10 has only 10 features)

under the window of RT <= 0.025min and *m*/*z* error <= 5 ppm (Figure 5). With the most stringent filtering (*m*/*z* and RT without errors), only the groups of $2x10^{-4}$ and $2x10^{-3}$ had more than 99% and 95% consensus features, respectively. When the window expands to *m*/*z* error <= 5ppm and RT shift <= 0.025min, the proportion of consensus features significantly increased and reached over 95% in all groups ($2x10^{-4}$ for 99.96%, $2x10^{-3}$ for 99.63%, $8x10^{-3}$ for 98.45%, $2x10^{-2}$ for 95.35%), except for $2x10^{-1}$ (62.78%).

In general, the error group of $2x10^{-1}$ could substantially change the feature list in the blank file. Eventually, the proportion of unique features is less than 1% of error threshold in two groups ($2x10^{-4}$, $2x10^{-3}$), which achieved a good performance when the original signal was processed for identification of features.

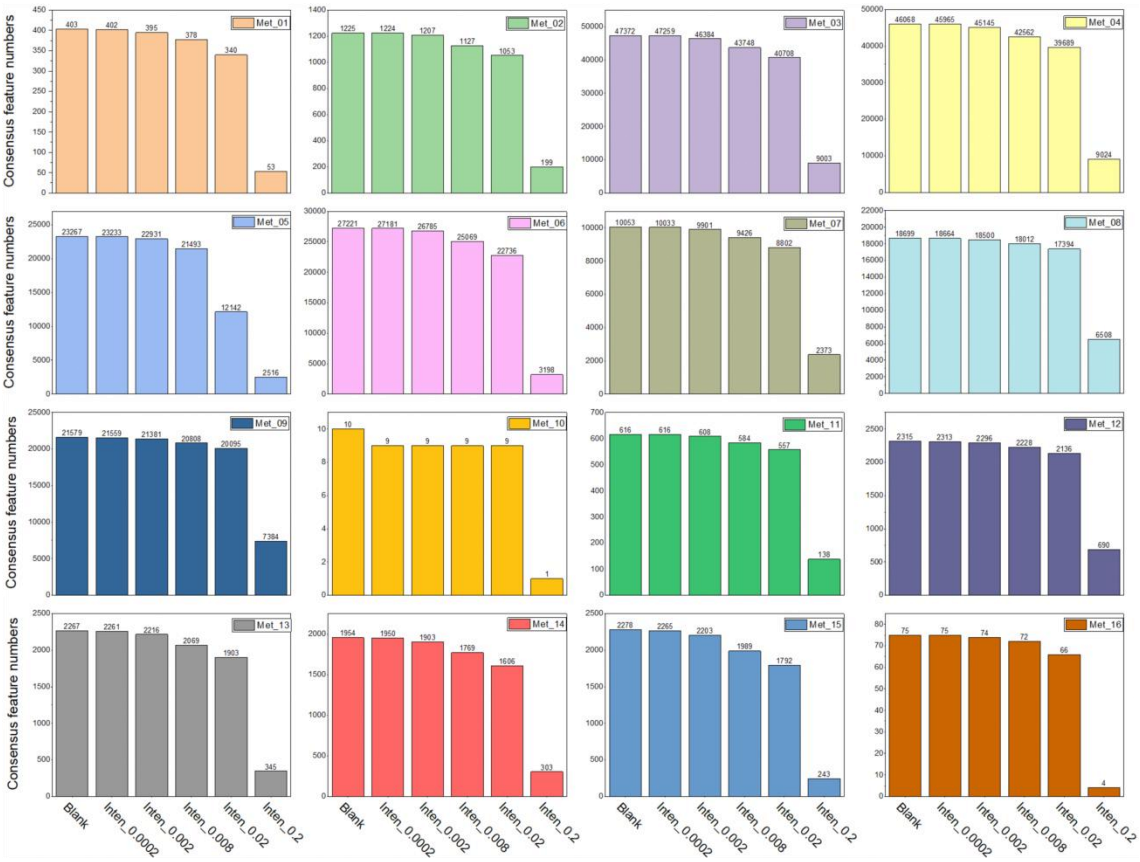


Figure 4. The number of consensus features detected in all groups (Blank, $2x10^{-4}$, $2x10^{-3}$, $8x10^{-3}$, $2x10^{-2}$, $2x10^{-1}$) across all files (Met_01 – Met_16) with truncated intensities.

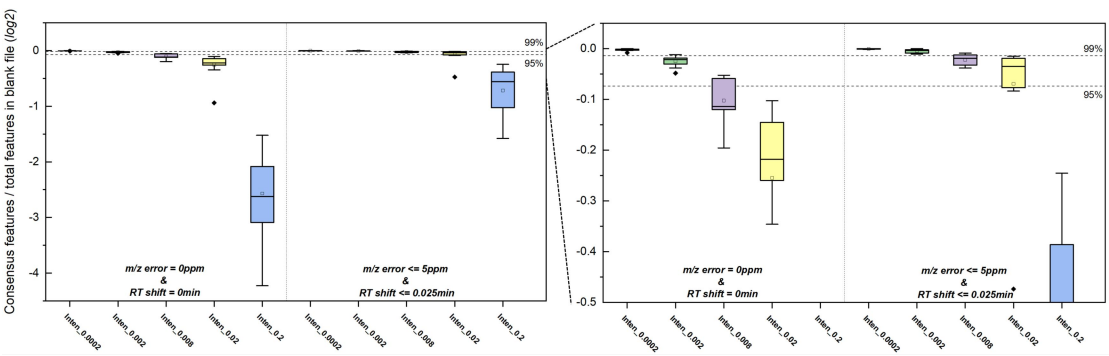


Figure 5. The ratio of consensus features to total features of blank files in each group with truncated intensities under different filtration windows (*m*/*z* & RT error equals 0, *m*/*z* <= 5ppm & RT <= 0.025min).

The ratio values were log_2 transformed. Met_10 was excluded due to the number of total features (n=10) was too small.

Consensus features shared the *m/z* and RT value, while the difference between their intensities was unknown and could impact the compound identification in the downstream of data analysis. For the filtered consensus features, we calculated the fold-change (FC) of their intensities between blank files and error groups (Figure6). We defined significant features with FC >2 (upper dash line in Figure6) or <0.5 (lower dash line in Figure6) and less significant features (0.5< FC <2, except FC=1). All groups have a small fraction of significant features (less than 0.5%), except about 0.63% (Met_12) and 0.55% (Met_06) in the group of $2x10^{-1}$. The proportion of less significant features in consensus features for groups were vary ($2x10^{-4}$ for <1%, $2x10^{-3}$ for <5%, $8x10^{-3}$ for about 20%, $2x10^{-2}$ for about 40%, $2x10^{-1}$ for about 95%). Therefore, the fraction of significant features was considered in a controllable level in all groups (mostly less than 0.5%). However, the fraction of less significant features suggested that $2x10^{-4}$ and $2x10^{-3}$ seems to be preferred error groups with a number below 5%.

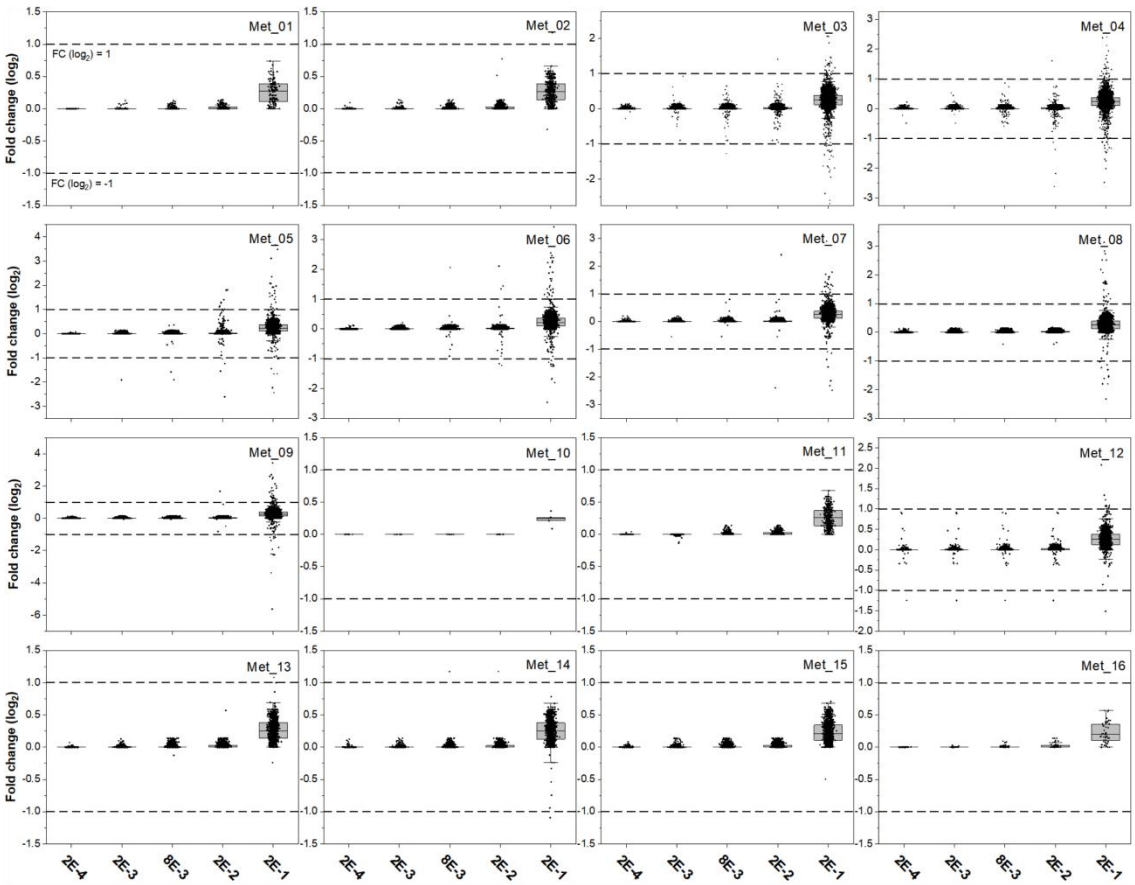


Figure 6. The summary of intensities fold-change (FC) of consensus features in all groups of files with truncated intensities. The box chart shows Log_2 transformed FC values in all groups ($2x10^{-4}$, $2x10^{-3}$, $8x10^{-3}$, $2x10^{-2}$, $2x10^{-1}$);

The differences between error groups and blank files could be caused during the data analysis process. In MS data processing studies, chromatogram construction of EIC and chromatographic peak detection are critical steps and are prone to generate false positive or false negative features [29]. Tracing back to modules used in our study, three of them (i.e., Mass detection, ADAP chromatogram builder, ADAP feature resolver) are likely related to the discovery of unique features [30]. In the mass detection module, due to these files being in

profile mode, the algorithm of Exact mass is selected for data centroiding. The errors of intensities might shift the location of data points closest to the peak center at half maximum intensity, which causes some errors in exact mass detection. For the module of ADAP chromatogram builder, two user-defined parameters (group intensity threshold and min highest intensity) might include some false EICs or exclude true EICs due to the changes of data points intensity values. In the ADAP peak detection module, we set the parameter of the Signal/Noise estimator to an intensity-based window method, where the smallest standard deviation of the intensities is used for the estimation of noises [30]. All intensity-related parameters that we used in MZmine3 might change the peak list to some extent due to the different relative errors in intensity values.

3.2.2 Comparison of Compound Lists among Intensity Groups with Different Relative Errors

The total number of detected compounds ranges from 3 (Met_10) to 9902(Met_04) (Table S3). After merging these compounds shared with the identical compound name, molecular formula, and neutral mass, only about half of them were kept, called "types". Three files (Met_10, Met_15 and Met_16) with a number of detected compounds less than 50 in all error groups were excluded from data statistics. The remaining 13 files contained a small fraction of unique compounds in the groups of $2x10^{-4}$ (average of 0.05%), $2x10^{-3}$ (average of 0.12%), $8x10^{-3}$ (average of 0.50%), $2x10^{-2}$ (average of 0.98%) (Figure 7). However, the groups of $2x10^{-1}$ observed a decrease in the number of consensus compounds and had a roughly 7% of unique compounds. Thus, a file with an intensity relative error of less than $2x10^{-2}$ could result in a variance of less than 1% in compound identification compared to the precision-lossless files.

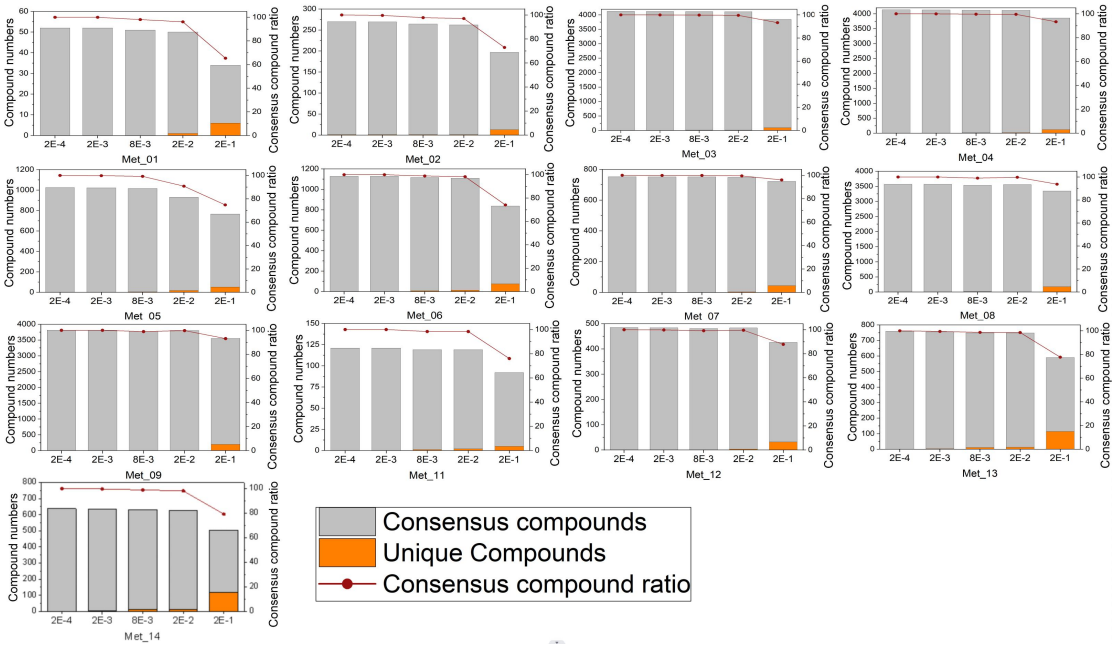


Figure 7. The number of total compounds and the percentage of consensus compounds in all five error groups across 13 mzML files with truncated intensity values.

3.3 Truncation Transform of m/z Values

Similar to intensities, *m/z* values with a defined error that would not affect the results of data processing have been evaluated by MSNumpress (0.002ppm), mspack (10^{-4} for absolute error, 0.2ppm for 500 *m/z*), and Stack-ZDPD (5dp, 0.1ppm) [12, 14, 15]. The latter two have similar results, but the MSNumpress showed a well lower relative error. Here, we used the absolute error truncation transform for *m/z* values. We set the first group with *m/z* under a

maximum absolute error of 10^{-5} (0.02ppm for 500 *m*/*z*), which lies between 0.2 ppm and 0.002 ppm. Based on the results, we further tested four degrees of error, $2x10^{-5}$, $5x10^{-5}$, 10^{-4} and 10^{-3}.

3.3.1 Comparison of Feature Lists among *m*/*z* Groups with Different Absolute Errors

The changes in the total numbers of detected features between blank and *m*/*z* error groups did not become significant until the error group (16 files) of $5x10^{-5}$ ($p < 0.05$) (Figure8 and Table S4). As the error increases, almost all consensus features are missing in the group of 10^{-3}. Next, we examined the proportion of consensus features in all detected features of a blank file and excluded the Met_10 file as the number of its detected peaks was too small to be statistically available (Figure9). With the hardest filtration criteria (*m*/*z* and RT shift equal zero), the group of 10^{-5} and $2x10^{-5}$ possessed the identical and slightly different (overlapping over 95% of consensus features) feature list as the blank file, respectively. However, the remain three groups ($5x10^{-5}$, 10^{-4}, 10^{-3}) had a fraction of consensus features from just below 95% to almost 0%. We then used the filtration window of *m*/*z* error <= 5ppm and RT shift <= 0.025min, resulting that consensus features took up over 99% of detected features in the first four groups (10^{-5} for 100%, $2x10^{-5}$ for 99.99%, $5x10^{-5}$ for 99.81%, 10^{-4} for 99.43%), but the group of 10^{-3} (92.39%) continued to lie below the line of 95%.

Overall, only the 10^{-3} error group changed the feature list by over 5% even after adjustment through the filtering window (*m*/*z* <= 5ppm and RT <= 0.025min). There are four groups (10^{-5}, $2x10^{-5}$, $5x10^{-5}$, 10^{-4}) under the error threshold (1%).

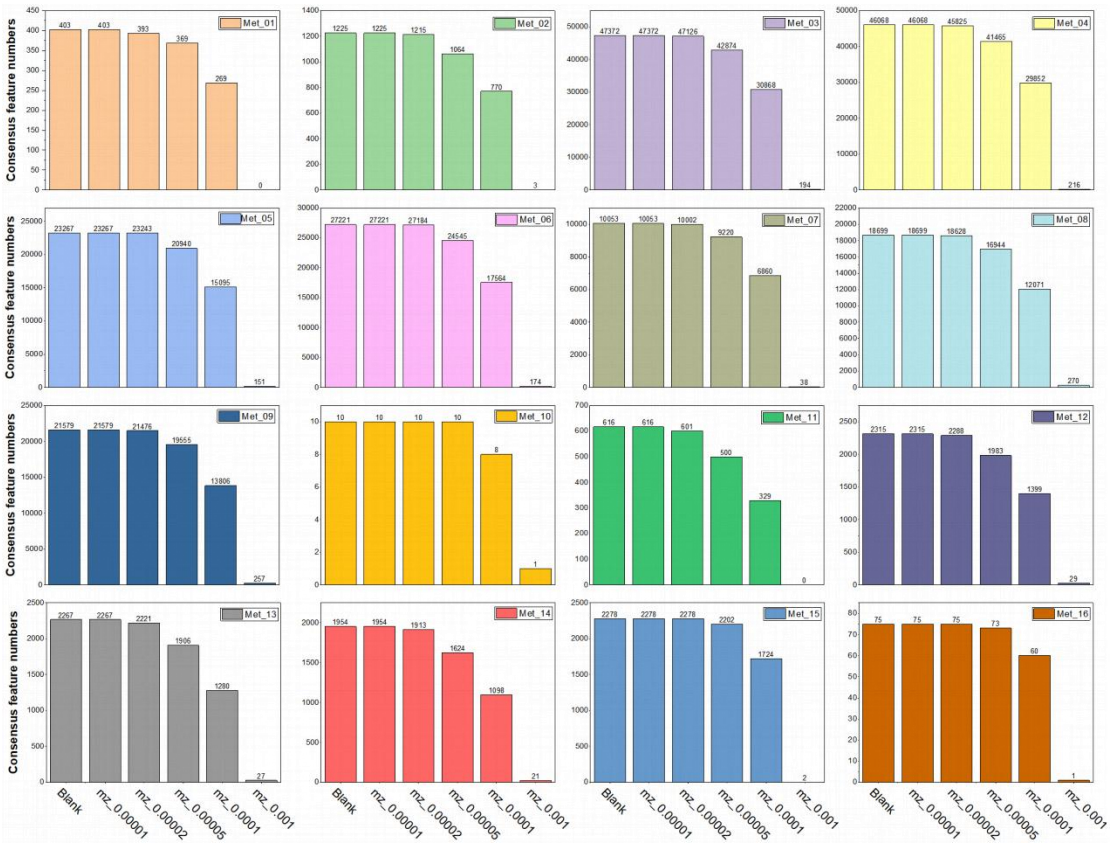


Figure 8. The number of consensus features detected in all groups (Blank, 10^{-5}, $2x10^{-5}$, $5x10^{-5}$, 10^{-4}, 10^{-3}) across all files (Met_01 – Met_16) with truncated *m*/*z* data.

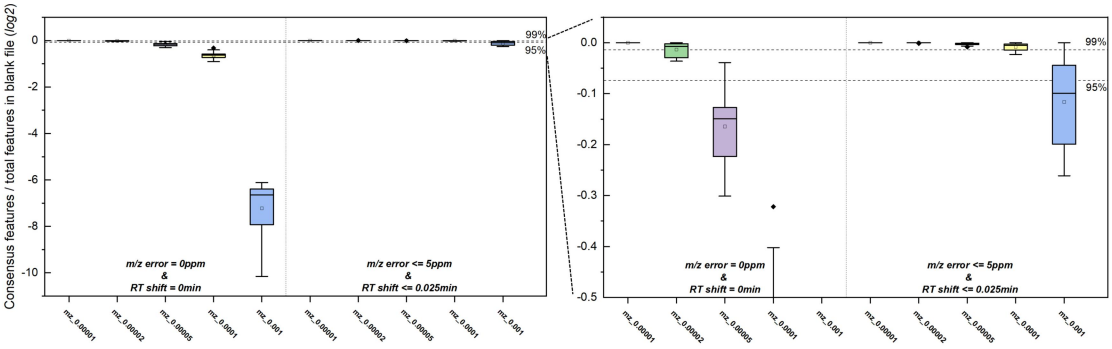


Figure 9. The ratio of consensus features to total features of blank files in each group with truncated *m/z* value under different filtration windows (*m/z* & RT error equals 0, *m/z* <= 5ppm & RT <= 0.025min). The ratio values were log_2 transformed. Met_10 was excluded due to the number of total features (n=10) was too small.

We further examined the intensity FC distribution of consensus features in all groups of files, and these significant features with FC > 2 or < 0.5 were outliers above the line (*log*2 value equals 1) or below the line (*log*2 value equals -1) (Figure10). For significant features, a small fraction of less than 0.5% occurs in all groups, except about 0.65% (Met_14) in the group of 10^{-2}. The less significant features were located between the two lines and represented a similar less than 5% of the number of detected features in all groups.

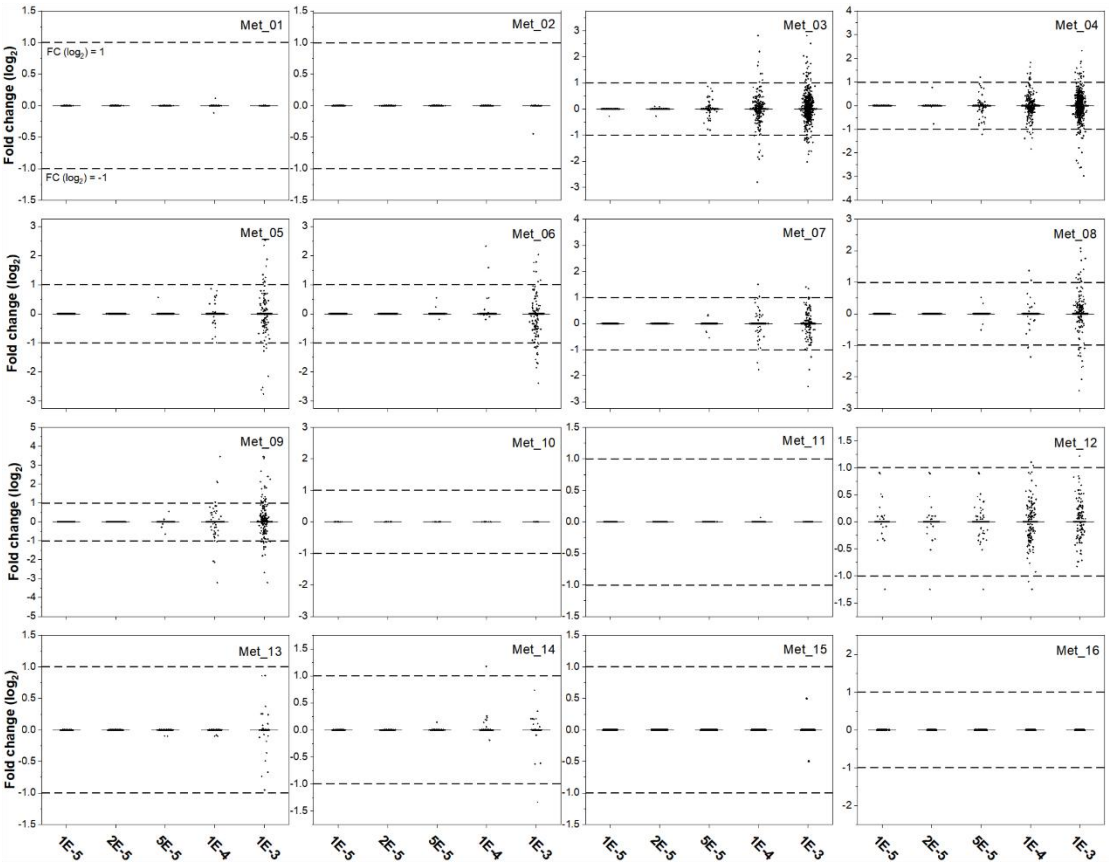


Figure 10. The summary of intensities fold-change (FC) of consensus features in all groups of files with truncated *m/z* data. The box chart shows Log_2 transformed FC values in all groups ($2x10^{-4}$, $2x10^{-3}$, $8x10^{-3}$, $2x10^{-2}$, $2x10^{-1}$).

To understand the variation generated from *m/z* error groups, tracking back to the data processing steps would be a preferred way. Features of misidentification continue to be a common issue for data processing algorithms and many attempts have been practiced minimizing its negative effects [28-32]. ADAP showed better performance in the number of good peaks with real peak shapes and boundaries than the traditional continuous wavelet transform (CWT) algorithm [28]. However, the robustness (the percentage of true features in total features) of ADAP illustrated a wide range from 6.79% to 87.57% across various types of samples [29]. This organism-dependent robustness may explain the fluctuations in fractions of true features across our data files. To discuss the effects of *m/z* values with errors on feature detection, we checked the ADAP peak-picking process again. For EIC building, all data points would be resorted from the highest intensity values to the lowest data and then be clustered to an EIC by the defined *m/z* accuracy (the parameter of scan-to-scan accuracy). In the study, *m/z* accuracy was set for TOF analyzers (0.005 *m/z* or 15 ppm) and orbitrap analyzers (0.005 *m/z* or 10 ppm) by the recommendation of MZmine3 documentation. The 15 ppm and 10 ppm at 500 *m/z* equal absolute errors of 0.0075 and 0.005 *m/z* (0.015 and 0.01 *m/z* for 15 and 10ppm at 1000 *m/z*), respectively. Therefore, the *m/z* with truncation transform at an absolute error of 10^{-3} is likely to be incorrectly included or excluded from the EIC-determining group. The errors from EICs would subsequently impact the peak detection step, where a certain part of misidentification features would not be filtered out as noise signals and eventually be kept as true features.

3.3.2 Comparison of Compound Lists among *m/z* Groups with Different Absolute Errors

All the feature lists generated from files (except for Met_10, Met_15 and Met_16) with *m/z* errors have been processed by the online database search module of MZmine3 with the same parameters we did in the intensity error study. The distribution of consensus and unique compounds showed in Figure 11. The number of unique compounds was increased by a noticeable degree from neglectable levels in the first four groups of 10^{-5} (0.03%), $2x10^{-5}$ (0.11%), $5x10^{-5}$ (0.52%), 10^{-4} (1.1%) to an average of 10.0% in the group of 10^{-3} (Table S4). Hence, the groups of 10^{-5}, $2x10^{-5}$, $5x10^{-5}$ and 10^{-4} could control the UCR to about and below 1% of error threshold.

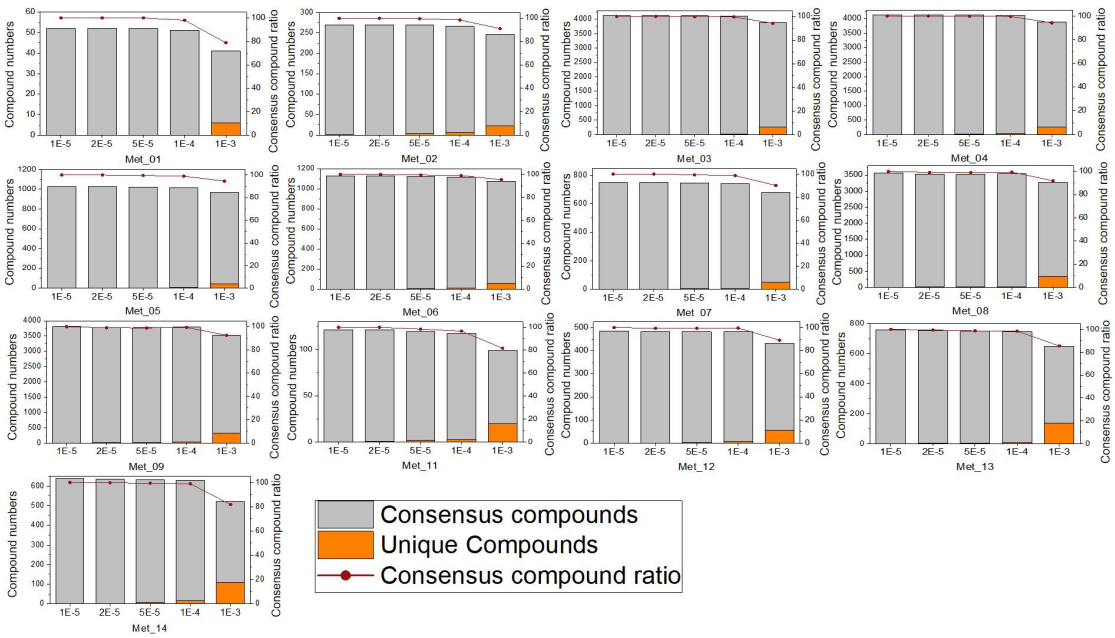


Figure 11. The number of total compounds and the percentage of cnsensus compounds in all five error groups across 13 mzML files with truncated *m/z* values

To conclude, we summarized the UFR, proportion of FC significant consensus features, and UCR in Table 12. For the intensity value, only two groups, $2x10^{-4}$ (0.04%) and $2x10^{-3}$(0.37%), had unique feature rates (UFR) of less than 1%, while the other groups had UFRs of varying degrees, such as $8x10^{-3}$ (1.55%), $2x10^{-2}$ (4.65%), and $2x10^{-1}$ (37.22%). We further checked the fraction of intensity FC significantly changed consensus features for each group, resulting in that all groups had a proportion of less than 1%. At the compound level, four of all groups ($2x10^{-4}$ for 0.05%, $2x10^{-3}$ for 0.12%, $8x10^{-3}$ for 0.50%, $2x10^{-2}$ for 0.98%, $2x10^{-1}$ for 7%) with a variation of <1%. Therefore, we consider that even though only errors less than $2x10^{-3}$ at the feature level can satisfy the condition of error threshold (1%), $2x10^{-2}$ after annotation of the feature list (i.e., compounds detection) is considered as a preferred truncation error for lossy compression of intensity values. For *m*/*z* value, there are three groups (10^{-5} for 0%, $2x10^{-5}$ for 0.01%, $5x10^{-5}$ for 0.19%, 10^{-4} for 0.57%) showed UFRs below 1% in feature detection and the other one (10^{-3} for 7.61%) had higher UFRs. Furthermore, the proportion of significant FC consensus features was below 0.5% in all groups. In compound detection, four of all groups (10^{-5} for 0.03%, $2x10^{-5}$ for 0.11%, $5x10^{-5}$ for 0.52%, 10^{-4} for 1.10%, 10^{-3} for 10%) with a UCR below or around 1%. Thus, we reckon that 10^{-4} can be used as a suitable truncation error for the *m*/*z* values, since it is below and closet to 1% of error threshold in both feature and compound levels. Taken together, we proposed a maximum absolute error of 10^{-4} for *m*/*z* and a maximum relative error of $2x10^{-2}$ for intensity.

Table 12. The summary of variations in all error groups in terms of unique feature rate, proportion of significant consensus features, and unique compound rate.

	Error Groups	Unique feature rate	Proportion of significant consensus features	Unique compound rate
intensity	0.0002	0.04%	<1%	0.05%
	0.002	0.37%		0.12%
	0.008	1.55%		0.50%
	0.02	4.65%		0.98%
	0.2	37.22%		7.00%
m/*z*	0.00001	0.00%	<0.5%	0.03%
	0.00002	0.01%		0.11%
	0.00005	0.19%		0.52%
	0.0001	0.57%		1.10%
	0.001	7.61%		10.00%

In this study, we only checked the untargeted metabolomics data in the MZmine3-based workflow, but the reproducibility of similar results on proteomics data and other software tools has not been examined yet. Moreover, the effects of precision loss on a downstream workflow of MZmine3, such as the feature based molecular network in GNPS [31], annotation in SIRIUS [32], need to be checked in our future step.

4. Conclusion

With the increase of MS data file size and volume, compressors specifically for MS data started to emerge. Among them, some lossy (near-lossless) compressors can reach a higher compression rate at the cost of precision losses for MS data (i.e., *m*/*z* and intensity). However, the extent to which the loss of precision at *m*/*z* and intensity have no negative impact on the subsequent data analysis has not been systematically discussed. In this study, sixteen untargeted metabolomics vendor MS files from different instruments, organisms, data-acquisition types, and file sizes, were used to evaluate the effects of files with truncation-transformed *m*/*z* or intensity data on the data processing (feature detection) and annotation (compound identification) steps via MZmine3. The open format mzML was used as import format.

Converting vendor files to mzML files via msConvert with default setting will encode *m*/*z* at 64-bit precision and intensity at 32-bit precision. First, we examined that whether this precision combination is optimal for the data processing. Each vendor file was then converted to four mzML files based on the combination of storage precision (32-bit or 64-bit) for *m*/*z* and intensity values. We identified that files with fixed *m*/*z* storage precision and different (32-bit and 64-bit) storage precision of intensity values tended to generate the identical feature lists in the MZmine3-based workflow. Files with fixed intensity accuracy and *m*/*z* data the different precisions generated different results in feature and compound detection, with an average variation of 0.13% for both feature and compound lists. We recognized that the difference of 0.13% is well below the error threshold (1%). Therefore, we chose mzML files (precision-lossless) with both *m*/*z* and intensity encoded in 32-bit precision for the further evaluation of precision losses.

A range of truncation transformations were applied to each precision-lossless mzML file, generating ten truncated (precision-lossy) mzML files, including five with only a maximum absolute error of *m*/*z* (10^{-5}, $2x10^{-5}$, $5x10^{-5}$, 10^{-4}, 10^{-3}) and five with only a maximum relative error of intensities ($2x10^{-4}$, $2x10^{-3}$, $8x10^{-3}$, $2x10^{-2}$, $2x10^{-1}$). We compared the results of feature and compound identification of precision-lossy files to precision-lossless files. A window of *m*/*z* error <= 5ppm and RT shift <= 0.025min was selected and used for consensus feature filtration. The UFR, FC significant consensus features, and UCR were calculated for all error groups of *m*/*z* and intensity data. We eventually presented a maximum absolute error of 10^{-4} for *m*/*z* and a maximum relative error of $2x10^{-2}$ for intensity, meeting the requirements of error threshold (1%). We further compared our recommendations to the recommended errors in existing near-losses compressors, such as MSNumpress ($2x10^{-9}$ for *m*/*z* and $2x10^{-4}$ for intensity) and mspack (10^{-4} for *m*/*z* and 10^{-2} for intensity). According to our results, we found that all error groups mentioned above would keep the variation of compound detection below 1%. However, we further extended the maximum relative error of intensity values to $2x10^{-2}$. We hope all results generated from our study could help researchers to make improvements to lossy compression algorithms in order to minimize the negative effects on downstream data processing, annotation and statistical steps.

Funding statement

This work was supported by the Shandong Provincial Natural Science Fund (2022HWYQ-081).

Author contributions

Tong Junjie: Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing – Original Draft, Writing – Review & Editing, Visualization. **Lu Miaoshan:**

Conceptualization, Methodology, Writing – Review & Editing. **Peng Bichen:** Validation, Writing – Review & Editing. **An Shaowei:** Investigation, Writing – Review & Editing. Wang Jinyin: Writing – Review & Editing. **Yu Changbin:** Conceptualization, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data included in article referenced in article.

Appendix A. Supplementary data

Supplementary data to this article can be found online at

References

1. Cui, L., H. Lu, and Y.H.J.M.s.r. Lee, *Challenges and emergent solutions for LC-MS/MS based untargeted metabolomics in diseases.* 2018. **37**(6): p. 772-792.
2. Chen, Y., E.-M. Li, and L.-Y.J.M. Xu, *Guide to Metabolomics Analysis: A Bioinformatics Workflow.* 2022. **12**(4): p. 357.
3. Schrimpe-Rutledge, A.C., et al., *Untargeted metabolomics strategies—challenges and emerging directions.* 2016. **27**(12): p. 1897-1905.
4. Blazenovic, I., et al., *Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics.* Metabolites, 2018. **8**(2).
5. Pluskal, T., et al., *MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.* BMC Bioinformatics, 2010. **11**: p. 395.
6. Schmid, R., et al., *Integrative analysis of multimodal mass spectrometry data in MZmine 3.* Nat Biotechnol, 2023.
7. Tsugawa, H., et al., *MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis.* Nat Methods, 2015. **12**(6): p. 523-6.
8. Tautenhahn, R., et al., *XCMS Online: a web-based platform to process untargeted metabolomic data.* Anal Chem, 2012. **84**(11): p. 5035-9.
9. Deutsch, E.W.J.M. and c. proteomics, *File formats commonly used in mass spectrometry proteomics.* 2012. **11**(12): p. 1612-1621.
10. Martens, L., et al., *mzML--a community standard for mass spectrometry data.* Mol Cell Proteomics, 2011. **10**(1): p. R110 000133.
11. Yang, R., X. Chen, and I. Ochoa, *MassComp, a lossless compressor for mass spectrometry data.* BMC Bioinformatics, 2019. **20**(1): p. 368.
12. Teleman, J., et al., *Numerical compression schemes for proteomics mass spectrometry data.* Mol Cell Proteomics, 2014. **13**(6): p. 1537-42.
13. Lu, M., et al., *Aird: a computation-oriented mass spectrometry data format enables a higher compression ratio and less decoding time.* BMC Bioinformatics, 2022. **23**(1): p. 35.

14. Hanau, F., H. Rost, and I. Ochoa, *mspack: efficient lossless and lossy mass spectrometry data compression.* Bioinformatics, 2021.
15. Wang, J., et al., *StackZDPD: a novel encoding scheme for mass spectrometry data optimized for speed and compression ratio.* Sci Rep, 2022. **12**(1): p. 5384.
16. Chambers, M.C., et al., *A cross-platform toolkit for mass spectrometry and proteomics.* Nat Biotechnol, 2012. **30**(10): p. 918-20.
17. Alka, O., et al., *DIAMetAlyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics.* Nat Commun, 2022. **13**(1): p. 1347.
18. Li, Z., et al., *Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection.* Anal Chim Acta, 2018. **1029**: p. 50-57.
19. John Peter, A.T., M. Peter, and B. Kornmann, *Interorganelle lipid flux revealed by enzymatic mass tagging in vivo.* 2021: p. 2021.08.27.457935.
20. Suvarna, K., et al., *A Multi-omics Longitudinal Study Reveals Alteration of the Leukocyte Activation Pathway in COVID-19 Patients.* J Proteome Res, 2021. **20**(10): p. 4667-4680.
21. Yan, Y., et al., *Plasma lipidomics analysis reveals altered lipids signature in patients with osteonecrosis of the femoral head.* Metabolomics, 2022. **18**(2): p. 14.
22. Tada, I., et al., *Correlation-based deconvolution (CorrDec) to generate high-quality MS2 spectra from data-independent acquisition in multisample studies.* 2020. **92**(16): p. 11310-11317.
23. Flasch, M., et al., *Stable Isotope-Assisted Metabolomics for Deciphering Xenobiotic Metabolism in Mammalian Cell Culture.* ACS Chem Biol, 2020. **15**(4): p. 970-981.
24. Muhsen Ali, A., et al., *Metabolomic Profiling of Submaximal Exercise at a Standardised Relative Intensity in Healthy Adults.* 2016. **6**(1): p. 9.
25. Danne-Rasche, N., S. Rubenzucker, and R. Ahrends, *Uncovering the complexity of the yeast lipidome by means of nLC/NSI-MS/MS.* Anal Chim Acta, 2020. **1140**: p. 199-209.
26. Tautenhahn, R., C. Bottcher, and S. Neumann, *Highly sensitive feature detection for high resolution LC/MS.* BMC Bioinformatics, 2008. **9**: p. 504.
27. Li, Y. and L. Li, *Retention time shift analysis and correction in chemical isotope labeling liquid chromatography/mass spectrometry for metabolome analysis.* Rapid Commun Mass Spectrom, 2020. **34 Suppl 1**: p. e8643.
28. Yu, F., et al., *Mass measurement accuracy of the Orbitrap in intact proteome analysis.* Rapid Commun Mass Spectrom, 2016. **30**(12): p. 1391-7.
29. Myers, O.D., et al., *Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data.* Anal Chem, 2017. **89**(17): p. 8689-8695.
30. Myers, O.D., et al., *One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks.* Anal Chem, 2017. **89**(17): p. 8696-8703.
31. Nothias, L.F., et al., *Feature-based molecular networking in the GNPS analysis environment.* Nat Methods, 2020. **17**(9): p. 905-908.
32. Duhrkop, K., et al., *SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite*

structure information. Nat Methods, 2019. **16**(4): p. 299-302.