Impact of three different peak picking software tools on the quality of untargeted metabolomics data

Wartmann, Yannick ; Boxler, Martina I ; Kraemer, Thomas ; Steuer, Andrea Eva

DOI: https://doi.org/10.1016/j.jpba.2024.116302

Journal of Pharmaceutical and Biomedical Analysis 248 (2024) 116302

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.journals.elsevier.com/journal-of-pharmaceutical-and-biomedical-analysis

Impact of three different peak picking software tools on the quality of untargeted metabolomics data

Yannick Wartmann, Martina I. Boxler, Thomas Kraemer, Andrea E. Steuer *

Department of Forensic Pharmacology and Toxicology, Zurich Institute of Forensic Medicine,University of Zurich, Winterthurerstrasse 190/52, Zurich 8057, Switzerland

ARTICLE INFO

Keywords:
(Un)targeted metabolomics
LC-MS/MS
DDA
DIA
Comparison
MS-DIAL
MZmine

ABSTRACT

Data quality and control parameters are becoming more important in metabolomics. For peak picking, open-source or commercial solutions are used. Other publications consider different software solutions or data acquisition types for peak picking, a combination, including proposed and new quality parameters for the process of peak picking, does not exist. This study tries to examine the performance of three different software in terms of reproducibility and quality of their output while also considering new quality parameters to gain a better understanding of resulting feature lists in metabolomics data. We saw best recovery of spiked analytes in MS-DIAL. Reproducibility over multiple projects was good among all software. The total number of features found was consistent for DDA and full scan acquisition in MS-DIAL but full scan data leading to considerably more features in MZmine and Progenesis Qi. Feature linearity proved to be a good quality parameter. Features in MS-DIAL and MZmine, showed good linearity while Progenesis Qi produced large variation, especially in full scan data. Peak width proved to be a very powerful filtering criteria revealing many features in MZmine and Progenesis Qi to be of questionable peak width. Additionally, full scan data appears to produce a disproportionally higher number of short features. This parameter is not yet available in MS-DIAL. Finally, the manual classification of true positive features proved MS-DIAL to perform significantly better in DDA data (62 % true positive) than the two other software in either mode. We showed that currently popular solutions MS-DIAL and MZmine perform well in targeted analysis of spiked analytes as well as in classic untargeted analysis. The commercially available solution Progenesis Qi does not hold any advantage over the two in terms of quality parameters, of which we proposed peak width as a new parameter and showed that already proposed parameters such as feature linearity in samples of increasing concentration are advisable to use.

1. Introduction

Metabolomics, the analysis of endogenous metabolites, is one of the big -omic fields still increasing in popularity and application with a rise from around 1400 publications on PubMed in the year of 2010 to above 12,000 in the year of 2022 [1,2]. Over the last years, metabolomics has been applied in different research fields, including biomarker research [3], drug discovery [4], the study of diseases [5], and very recently, forensic medicine or clinical and forensic toxicology [6–9].

In principle, the investigation of metabolome changes can be tackled by two different analytical approaches – targeted or untargeted [3, 10–14]. While targeted analysis evaluates changes in a preselected, limited number of (endogenous) metabolites, untargeted metabolomics tries to analytically capture theoretically all metabolite information within a sample. This is followed by a selection of statistically relevant features in later stages. Untargeted metabolomics hugely benefitted from advancements in measuring techniques but even more so from an increase in available computational power and improved software for the analysis of untargeted data. The acquisition and particularly processing of untargeted metabolomics data is a long and multi-stepped process also associated with various pitfalls. When data has been acquired, one of the first steps after data conversion (if necessary) would be to perform an automated peak picking. Further data processing steps include data cleaning, normalization, and transformation before statistical evaluation of interesting features and their annotation. Numerous different commercial, freeware, or open-source software (packages) are nowadays available for a whole untargeted metabolome workflow or different sub-steps of it [15,16]. Overall, comparative analysis of the

* Correspondence to: University of Zurich, Zurich Institute of Forensic Medicine (ZIFM), Department of Forensic Pharmacology & Toxicology, Winterthurerstrasse 190/52, Zurich CH-8057, Switzerland.
E-mail address: andrea.steuer@irm.uzh.ch (A.E. Steuer).

https://doi.org/10.1016/j.jpba.2024.116302
Received 27 March 2024; Received in revised form 7 June 2024; Accepted 8 June 2024
Available online 10 June 2024

influence of the different parameters and/or different software solutions on the outcome of an untargeted metabolome experiment is scarce [17]. However, it is evident that each of these steps can introduce changes and variations to the data set and as such it is not surprising that depending on the chosen analytical settings (sample preparation, analytical methods, etc.) and processing (software, peak picking, data normalization or transformation, statistics, etc.) results might differ. [18] Even more so, there seems to be a lack of measurable quality markers in processed data [17,19]. Overall, little data exists regarding the reproducibility of findings obtained in untargeted metabolome analysis.

The topic of different data acquisition modes, mainly comparing Full Scan, data-independent (DIA), and data-dependent acquisition (DDA), and various software options was partially explored by Guo *et al.* [19] and others [20–23]. Regarding different software tools and their influence on the results, Li et al. evaluated a mixture of 1100 analytes utilizing four different software solutions. 130 of those analytes significantly varied in concentration. They proposed the linearity of concentration increase across samples as a quality marker, something others have proposed as well [24–27]. Hemmer et al. compared Compound Discoverer 3.1, XCMS online, and a custom R script for their suitability in untargeted metabolomics data analysis. The group relied on an established measuring procedure of full scan acquisition followed by the acquisition of only MS/MS spectra of significant features. While this might ensure high-quality MS/MS data, it involves multiple measurements and hence increases workload [18]. Liao *et al.* have studied the influence of Progenesis Qi, MarkerView, and XCMS on metabolite annotation and pathway enrichment while considering signal drift and peak number as parameters [28]. Other groups have made suggestions about parameter settings and algorithm improvements in certain software [29–31]. A common problem to all workflows remains feature quality and the detection of true positive features. Analyzing the above-mentioned publications and our own work, it becomes apparent that the total number of features detected by a certain software does not indicate quality or hint at the rate of true positive features in any way. It also shows that there is huge variation in the utility of certain software for differently acquired data, e.g., different measuring set-ups such as gas- or liquid chromatography and different mass analyzers.

Following the lack of sufficient comparative studies on different software tools and quality markers for data processing the aim of the present study was to compare currently employed software solutions offering a graphical user interface (GUI) and the possibility of processing metabolomics data in a full workflow consisting of data loading, peak picking, alignment, gap filling and if needed annotation and/or library search. For the chosen software tools, little to no programming knowledge should be required as the software should be easy to deploy in any lab environment. Therefore, MZmine 3 [45] and MS-DIAL (msDial) 4.9 [46], as well as the commercially available vendor software Progenesis Qi (ProgQi), should be compared with regard to their peak picking capabilities focusing on LC-MS data acquired in DDA as well as FullScan mode.

2. Materials and methods

2.1. Materials

Adenine, adenosine, alanine, arginine, butyrylcarnitine, caffeine, cholic acid, creatinine, decanoylcarnitine, dodecanoylcarnitine, dopamine, glycocholic acid, hexadecanoylcarnitine, hippuric acid, histidine, inosine, isoleucine, kynurenine, leucine, lysine, methionine, octadecanoylcarnitine, octanoylcarnitine, ornithine, phenylalanine, proline, propionylcarnitine, taurine, taurocholic acid, tetradecanoylcarnitine, theobromine, theophylline, threonine, tryptophane, tyrosine, uridine, ursodexcholic acid and were purchased from Sigma Aldrich (Buchs, Switzerland). Deuterated and heavy labelled internal standards (IS) adenosine ribose-D1, arginine-13C6, caffeine 3-methyl-13 C, carnitine trimethyl-D9, creatinine N-methyl-D3, deoxycholic acid-D4, dopamine-D4, glycine-13C2, glycocholic acid-D4, hippuric acid 15 N, kynurenine-D4, leucine-D10, lysine-D4, phenylalanine-D1, proline 15 N, serine-D3 and tryptophan-D5, were purchased from Cambridge isotope laboratories, which were delivered by ReseaChem Life Science (Burgdorf, Switzerland) or Sigma Aldrich (Buchs, Switzerland). Water, acetonitrile (ACN), methanol (MeOH) of HPLC grade were obtained from Fluka (Buchs, Switzerland). All other chemicals used were from Merck (Zug, Switzerland) and of the highest grade available.

2.2. Sample preparation and measurement

QC samples were prepared and measured as described previously by Boxler *et al.* [23]. In summary, authentic plasma was collected from a healthy volunteer and stored in ammonium heparin tubes at −20 °C. The volunteer has given written informed consent. Protein precipitation was performed using methanol. The plasma was once analyzed in its native form (QC1) and additionally fortified with known analytes (n=38 analyzed in the current study) in increasing concentrations (QC2 lowest concentration, QC4 highest concentration; chosen analytes and concentrations given in Supplementary material Table S1 [23]) on four different days of which samples from one day were chosen for this study. Measurements were performed in duplicate (Sample A and B) for each QC level. All samples were randomized and measured on a Thermo Fischer Ultimate 3000 UHPLC system (Thermo Fischer Scientific, San Jose, CA, USA) coupled to a high-resolution TOF instrument system (TripleTOF 6600 Sciex, Concord, Ontario, Canada). The mobile phases A and B consisted of 10 mM ammonium formate and 0.1 % (v/v) formic acid in water or 0.1 % (v/v) formic acid in methanol respectively. The liquid chromatography was performed using a reversed phase (RP) column (XSelect HSST RP-C18 column; 150 mm×2.1 mm i.d; 2.5 μm particle size; Waters, Baden, Daettwil, Switzerland). The gradient was set as follows: 1 minute 100 % A; 1–15 minutes 100 % B; 15–18 minutes held at 100 % B and then decreased to starting conditions and re-equilibration for 2 minutes. The flow rate was increased after 15 minutes to 0.7 mL/min. MS and MS/MS data was acquired by two methods: TOF-MS only (no MS/MS information acquired, FullScan) and data dependent acquisition (DDA). The analysis was performed using a DuoSpray ion source at a resolving power (full width at half maximum at m/z 400) of 30,000 (high resolution mode) or 15,000 (high sensitivity mode) for MS and MS2 in positive electrospray ionization (ESI) mode. The FullScan method scanned over a mass range of m/z 50 to m/z 1000 with an accumulation time of 100 msec at 5 eV collision energy (CE). The DDA method included a FULLSCAN scan over the same mass range but with 50 msec accumulation time at a CE of 5 eV. Subsequent DDA experiments, each experiment with an accumulation time of 100 msec at a CE of 35 eV with a spread of 15 eV, were preformed after dynamic background subtraction with the four most intense ions above 100 counts per second (cps) and an exclusion time of 5 s after two occurrences in high sensitivity mode.

2.3. Data processing

This study used data from the FullScan only and from the DDA measurements. Each set of data (FullScan or DDA) was analyzed in **ProgQi** (Version 2.4), **MZmine 3** (Version 3.3.0), and **msDial** (Version 4.9.22),

Parameters in each software were optimized as much as possible (based on the recovery rate of spiked/known analytes and total features found) while keeping parameters available in more than one software identical (e.g., m/z tolerances available in msDial and MZmine were kept at m/z 0.005, while the noise filter was only available in MZmine and hence optimized for best results). The parameter ranges tested for optimization, as well as the final parameters chosen are given in the supplementary information Table S2, grouped by the respective workflow step they belong to.

Data processing in the three different software tools was performed

in two different subsets:

- A: "Targeted" data analysis for selected analytes: For data analysis concerning spiked analytes, each QC sample pair (e.g., QC1 sample A and QC1 sample B) was analyzed in a separate project per QC level and data acquisition mode (FullScan, DDA) using the parameter settings described above.,
- B: "Targeted" and untargeted data analysis for total features: To analyze the total number of features found by each program, all QC sample levels were analyzed together in one project per acquisition condition (FullScan or DDA).

After data processing in each software, the alignment results were exported as text file (either.csv or.txt) and imported into R (Version 4.2.3). Further data analysis was performed using R and R packages shown in Supplementary material Table S3.

2.4. Data evaluation

2.4.1. Identification capabilities for spiked analytes (targeted) and for all features (untargeted)

All features found in the individual QC datasets (data evaluation A and B) were compared to the expected data of the spiked reference standards with an allowed error margin of 0.005 *m/z* and 0.3 min retention time difference. If multiple features matched these criteria, the one with the closest retention time to the reference was chosen to be the correct one. To allow for comparability between all methods and processing variants, *m/z* and retention time were the only identifying metrics for features found.

In addition, the number of total features without additional filter criteria was evaluated.

2.4.2. Linearity of target compounds

Linearity of the 38 target compounds was assessed with a weighted (1/x) linear model and its corresponding R^2. All 38 models were manually confirmed to exclude the chance of arbitrarily good R^2 results.

The mean percentage difference (over all 4 QC levels and duplicates per level) of peak areas determined by the software, to manually curated peaks, was calculated. Per level, percentage difference was then subtracted by the above-described mean percentage difference.

Potential quality indication parameters: peak width

For MZmine and ProgQi, all detected peaks were divided according to their peak widths in too short (<0.05 min), adequate (> 0.05 min, < 0.50 min), and potentially too long peak width (stratified in >0.50/1.00/2.00 min categories in the context of the here employed chromatography) following a filter step to exclude peaks with a peak area of less than 300 cps.

2.4.3. Manual peak classification

Following a sample size calculation (N=10,000, z=1.96, $\hat{p}$=0.5, ε=0.07), 200 features of each dataset (no filtering) containing all QC samples (data evaluation B, DDA, FullScan, all 3 software) were manually selected at random. These 200 features were then analyzed and based on expert opinion, classified in either true positive or false positive. Criteria included peak shape, intensity, resolution, overlap between QC samples of the same level and overlap or linear increase between QC samples of increasing concentrations (in case the feature was part of one of the spiked analytes).

2.5. Software performance

Each software has been additionally evaluated for its performance. For this, our data set of 8 samples (duplicates per QC level, one sample set per acquisition mode) has been processed in each software as described above (processing method B). A timer was started as soon as the software began its independent data processing but only after entering all processing parameters. The timer was stopped once the software returned a final feature list that could be exported. This ensures that only raw software performance is reported and no user bias (entering, parameters, opening software, finding samples etc.) is included. Because ProgQi works slightly different compared to the other two, we report both the time for its automatic processing as well as the first step which in ProgQi is creating the experiment. All performance analysis was carried out on the same workstation equipped with dual Intel Xeon E5–2630 v4, 128 GB of ECC DDR4 memory running at 2133 MHz and an ATA ST2000DM001–1ER1 SCSI hard drive with a capacity of 1.8 TB.

3. Results

3.1. Parameter optimization

In ProgQi the peak picking sensitivity was tested at 3, 4 and 5. With the highest number of spiked analytes recovered at sensitivity 5, this setting was chosen. For the other software, with parameters such as mass accuracy in MS1, scan-to-scan accuracy, noise level, peak duration range, and minimum # of data points, an estimation from manual data inspection was made and compared to sensible presets recommended by the developers. In all cases, these were in line with each other. Testing extreme values such as a massively lower mass accuracy (0.02, 0.05) or very tight retention time allowance (0.05, 0.1) proved to be less effective at analyte recovery. Minimum peak height was tested in msDial at various levels (200, 300, 500, 1000) with the last setting recovering the most analytes while giving an adequate number of total features. Settings available in multiple software (such as minimum peak height) were kept identical to ensure similar starting conditions. Deviating from sensible presets recommended by the developers in all other parameters did not lead to significant improvements for spiked analyte recovery. Alignment parameters only mattered for the untargeted analysis of this study (processing B), as all other analyses were done in separate projects with only one QC level per project. Correspondingly, we found that even with separate projects, the retention time of known analytes in the targeted evaluation did shift minimally, with most exhibiting differences between projects (and hence QC levels) of less than 0.1 min. Depending on the software, few analytes shifted up to 0.2 (most in ProgQi) and only one analyte in ProgQi shifted between 0.2 and 0.3 min.

3.2. Identification capabilities for spiked analytes (targeted)

In Table 1 the number of analytes found after automatic peak picking with the different software tools is shown. Two different data processing strategies were applied, separate for each QC level (data processing A) and all QC levels combined (data processing B). Although not done in a real untargeted metabolome experiment, data processing A was performed not to influence peak picking, alignment and especially gap filling by higher concentrated samples. Recovery of individual analytes on all QC levels are shown in Supplementary material Table S4.

msDial found the most spiked analytes at every level (processing A),

Table 1
Recovery of known analytes: Number of features found per QC level (data processing A) and the number of features found at all QC levels (complete cases, data processing B) for each software and acquisition mode. Manually, a total of 38 analytes was identified in the dataset.

	msDial		MZmine		ProgQi	
QC Levels	DDA	FullScan	DDA	FullScan	DDA	FullScan
1	34	35	31	31	26	28
2	34	35	34	34	31	31
3	36	35	33	34	29	32
4	36	36	35	35	32	31
complete features	30	34	31	31	23	27

only being tied by MZmine at QC level 2 in DDA mode. A total of 38 analytes were found when manually identifying peaks in the dataset. With 36 analytes found at QC level 3 in the DDA dataset and QC level 4 in both DDA and FullScan data, msDial comes closest to this number. Additionally, the data is summarized over all QC levels (processing A) and only spiked analytes are considered that were found on all QC levels. In this analysis msDial finds the most analytes but only in FullScan data. In DDA acquisition, MZmine shows best recovery with ProgQi in last place.

3.3. Peak picking capabilities for all features (untargeted)

Analyzing data originating from the datasets processed with all samples of one acquisition condition at once (hence allowing for gap filling, data processing B) the total numbers of found features are shown in Table S6. Additionally, the data was filtered for features found on all QC levels. These features are called "complete features". There is virtually no difference in numbers between DDA and FullScan, and total vs. complete features for msDial, indicated by of the near 100 % retained features after filtering. A difference between DDA and FullScan data is seen in MZmine with more than twice as many detected following FullScan acquisition (18,851/18,412 DDA, 55,601/55,183 FullScan). ProgQi shows the largest discrepancy between feature numbers from DDA and FullScan (5568/4898, 88 %, DDA; 35,628/30,482, 86 % FullScan) with similarly lower percentage of retained features for both acquisition types compared to msDial or MZmine.

3.4. Potential quality indication parameters

3.4.1. Linearity of target compounds

All target analytes were spiked in increasing concentrations, the observed linearity is, exemplified for adenosine, shown in Fig. 1. Both msDial and MZmine show a comparable linear model for DDA and FullScan data, with MZmine having lower and msDial higher areas than the manually curated peaks. ProgQi shows a much bigger difference in slope between DDA and FullScan data. Results of all linear models are shown in Supplementary material Table S5.

The mean percentage difference (over all 4 QC levels and duplicates per level) of peak areas determined by the software, to manually curated peaks, was calculated. Per level percentage difference was then subtracted by the above-described mean percentage difference. Results of this calculation are shown in Fig. 2. In this analysis MZmine takes a clear lead with the least variability, meaning it showed the biggest consistency in peak area calculation. This nearly perfect consistency drops slightly off in FullScan data with a minor upward trend towards higher concentrated samples. These results are followed by msDial's performance. Results for msDial are influenced by a few analytes having higher variability, while many remain constant. There seems to be no clear trend for either acquisition type. ProgQi takes last place with strong variability, especially in FullScan data.

3.4.2. Peak width, area, and complete cases

To find further potential quality-indicating parameters, the peak width and area of features was looked at. Unfortunately, only MZmine and ProgQi offer both parameters in their output. The author could at the time of writing (msDial, version 4.92) not find a way to export peak width information with msDial or any other parameters from which peak width could have been back calculated. Table 2 shows the number of features of short (<0.05 min), adequate (>0.05 min and < 1.00 min) and longer peak width (> 1.00 min, stratified) with a minimum peak area of 300. Except for the few spiked analytes, most features should be present in all QC samples, hence the number of features found in all samples (complete features) is also looked at.

3.5. Manual peak classification

An overview of a manual peak classification of 200 randomly selected features as "true positive" peaks vs. noise or insufficient peaks (false positive) for integration is given in Fig. 3. Overall, only slightly more than half of the peaks were considered true positives, with msDial providing the best result in DDA (62 %). FullScan data generally showed worse performance than DDA. When calculating the absolute number of true positive features using the respectively found "incidence" and an unfiltered dataset, FullScan outperforms DDA when using MZmine (FullScan: 11,051true pos. features; DDA: 5678true pos. features) and ProgQi (FullScan: 2493 true pos. features; DDA: 1726 true pos. features) but not in msDial (FullScan: 3192 true pos. features; DDA: 6553 true pos. features).

3.6. Software performance

As shown in Table 3, MZmine was superior to the other two software in terms of processing time. While MZmine cannot handle native.wiff files, msDial can but benefits massively from conversion in.mzML as seen in the timings. ProgQi does not seem to be influenced by file conversion. In fact, it performed slightly worse when processing the open.

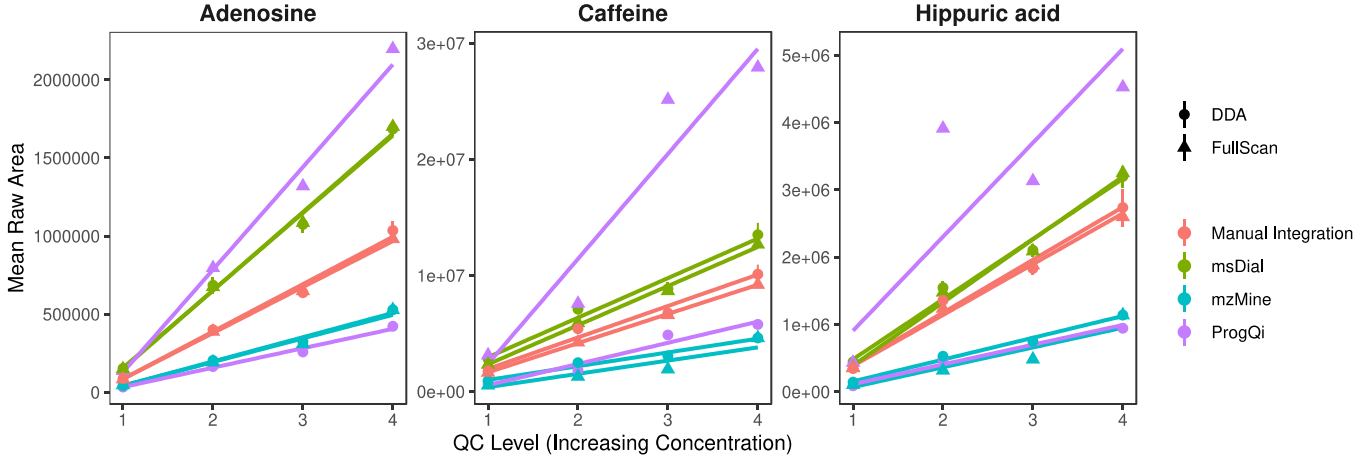

Fig. 1,. Linearity of known analytes: Examples of linearity (adenosine, caffeine, hippuric acid) over 4 QC levels (x-axis; individual concentrations per analyte and QC level are given in Table S1) in each software and data acquisition mode are shown. Except for ProgQi, each software or manual integration observes high similarity between DDA and FullScan data. There is no clear explanation for the huge discrepancy of DDA and FullScan data in ProgQi.

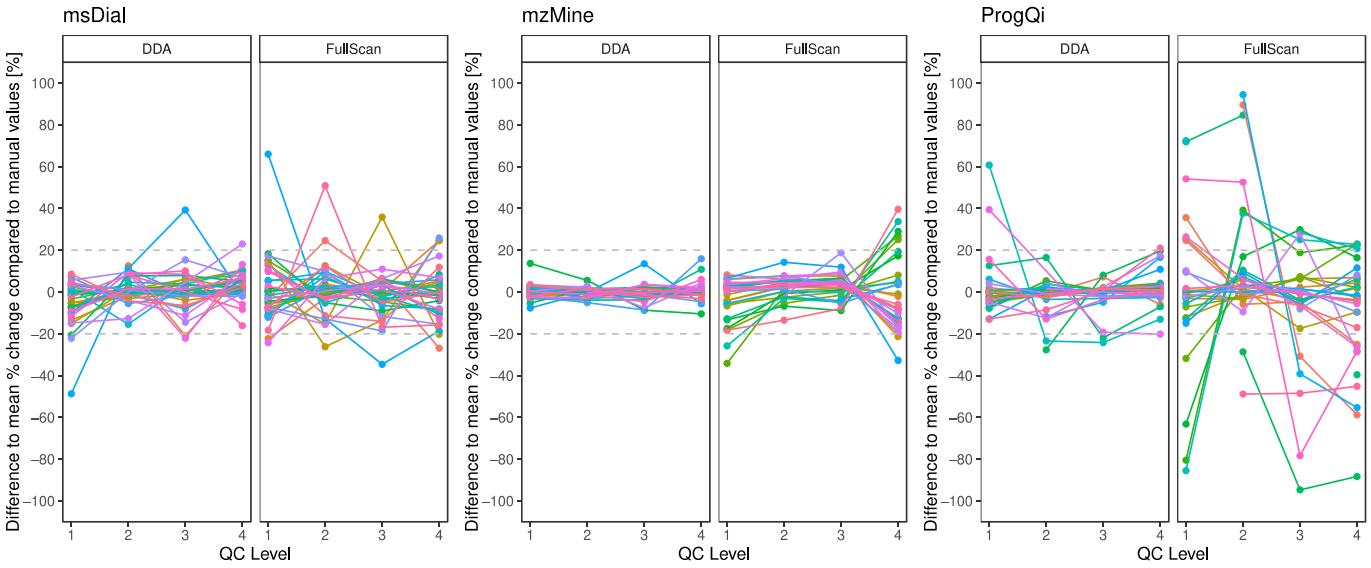

Fig. 2. Variability of difference to mean: The variation of percentage change per level compared to manual peak areas is shown for all three software and both data acquisition modes over 4 QC levels (x-axis, individual concentrations per analyte and QC level are given in Table S1). While MZmine shows very little variation, closely followed by msDial, ProgQi shows large variation.

Table 2
Feature width subsets: Features with a peak area > 300 are shown. The retained percentage of features compared to no peak area filter is indicated as well. Features with a peak width shorter than 0.05 min are considered too short. Features with a peak width greater than 0.50 min could potentially be too long and are shown in stratified manner below the likely adequately wide features.

	msDial		MZmine		ProgQi	
	DDA	FullScan	DDA	FullScan	DDA	FullScan
Total features* (% retained after area filter)	10,570 (100 %)	10,643 (100 %)	18,851 (100 %)	55,256 (99 %)	5545 (99 %)	34,019 (95 %)
< 0.05 min	/	/	4237	44,560	1407	29,693
> 0.05 min, < 0.50 min	/	/	**14,500**	**10,630**	**3900**	**4051**
> 0.50 min, < 1.00 min	/	/	114	64	199	231
> 1.00 min, < 2.00 min	/	/	0	2	27	44
> 2.00 min	/	/	0	0	12	0
% adequate peaks	/	/	**76.92**	**19.24**	**70.33**	**11.91**
Complete features*	10,404	10,147	18,224	47,409	4638	26,454
< 0.05 min	/	/	4122	37,222	951	22,316
> 0.05 min, < 0.50 min	/	/	**13,991**	**10,122**	**3465**	**3870**
> 0.50 min, < 1.00 min	/	/	111	64	193	228
> 1.00 min, < 2.00 min	/	/	0	1	22	40
> 2.00 min	/	/	0	0	7	0
% adequate peaks	/	/	**76.77**	**21.35**	**74.71**	**14.63**

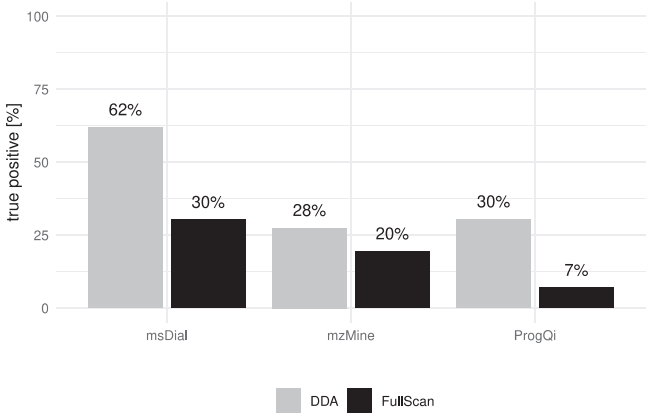

Fig. 3. True positive feature rate: The percentage of true positive features (n=200, chosen at random) is shown.

Table 3
Processing times of each software for the two sample sets (DDA/FullScan) for both vendor format files (.wiff) as well as converted files (.mzML). Time is reported as mm:ss. Total processing time for ProgQi is followed by the time it took to create the experiment and the then the automatic processing with the entry of parameters excluded from timings.

	msDial		MZmine		ProgQi	
File Format	DDA	FullScan	DDA	FullScan	DDA	FullScan
wiff	07:20	48:31	/	/	05:14 (01:46/ 03:27)	18:18 (01:49/ 16:29)
mzML	02:24	15:49	01:27	08:10	05:19 (01:29/ 03:50)	25:54 (01:38/ 24:16)

mzML format compared to the vendor specific.wiff file.

4. Discussion

Nowadays, numerous different commercial, freeware, or open-source software (packages) are available for a complete untargeted metabolome workflow. Awareness among users that depending on the chosen analytical settings (sample preparation, analytical methods, etc.) and processing (software, peak picking, data normalization or transformation, statistics, etc.), results might differ seems to be widespread. Nonetheless, few systematic studies on the influence of the different software tools on processed data quality are available. We wanted to extend on the existing knowledge of software influences by comparing a commercially available software, ProgQi, to software that nowadays seems to be state of the art and was mentioned in an increasing number of studies (msDial, MZmine). The main rationale for including the two latter ones was the focus on software tools, that use a well-explained GUI and offer the possibility to process metabolomics data in a full workflow (peak picking, alignment, gap-filling, etc.) with a feature table for further evaluation as an output. Although a metabolomics study's final goal is to obtain said feature tables for further statistical analysis, potential model building, and/or biomarker identification, we focused on maximizing the potential for true positive hits during the initial data processing phase. Therefore, we chose spiked matrix samples to include and compare targeted and manual comparison of features/analytes as a measure of software performance. In an authentic metabolome study, downstream statistical processing would be required, which is highly individual to a particular study's purpose, something we deliberately excluded from this study. However, other groups have presented various options for downstream processing such as Ramell *et al.* [32]. Often, the number of picked features, most likely due to the lack of alternative quality markers and its simplicity, is considered a or even the major criterium when optimizing data processing workflows. As such, studies preferred FullScan acquisition followed by separate acquisition of MS/MS generating methods as superior to solely DDA acquisition [18, 33]. Our study focused on more detailed criteria, elaborating on previous works [18,19,27,30,34,35], including not only the number of picked features (untargeted evaluation) but also the recovery of known present analytes in varying concentrations (targeted evaluation), the linearity (targeted), peak width and peak area (untargeted), and to the best of our knowledge for the first time manual inspection of true positives in a representative number of features.

Data used in this study was previously acquired for LC-MS method evaluation by Boxler *et al.* [23] The selected analytes were chosen based on identifications in previous forensic and other metabolome studies. The decision to choose only data acquired on RP (ESI positive) in FullScan and DDA mode was based on the fact that for software comparison a dataset had to be selected that was supported by all software packages (e.g., DIA – SWATH was not supported in our version of ProgQi) and that was not introducing further complications and potential pitfalls in analysis (e.g., combining multiple measuring mode results into one dataset for analysis). This selection reduced the number of initially spiked analytes to 38 manually, reliably detected analytes chosen as target compounds in the current data analysis comparison. [23] Data used in this study originates from spiked authentic plasma samples. It is important to consider potential "matrix effects" on the performance of the software. While plasma still is one of the most widely used matrices for metabolomic studies with the increase in sensitivity of untargeted mass spectrometry, less concentrated matrices such as oral fluid are examined as well. It cannot be guaranteed that our findings hold true for such low abundant and concentrated or even higher concentrated matrices (e.g. urine) as well. Additionally, while not actively observed in this study, the spiking and, hence, increase in the concentration of certain analytes (even above common concentration ranges) could influence our findings through well-known mechanics such as signal enhancement/suppression, among others. Nonetheless, the data used in this study can be considered representative of a standard metabolomics study using plasma samples.

Regarding the number of identified target compounds (Table 1), the performance of all three software is very close. While msDial seems to have a slight edge over the others, the difference is mostly around 10 % to MZmine or ProgQi. Exceptions are observed at lower QC levels, e.g., at QC1 where ProgQi finds 26 or 28 analytes in DDA or FullScan data respectively while msDial already recovers 34 or 35 analytes. QC3 shows an unexplained slight decrease in recovery rate for both MZmine and ProgQi but not for msDial. We have no clear explanation for this observation at the time of writing. None of the software managed to match the manual peak identification where a maximum of 38 analytes were confidently identified. This is somewhat explained by the fact that analytes not found in all or at least most software and acquisition modes were either low in intensity or already difficult to integrate manually. These included: dopamine, kynurenine, or taurocholic acid, among others. A complete list can be found in the Supplementary material Table S4. When limiting our dataset to analytes that were only found at all QC levels, hence also in all 4 separately performed peak picking projects, only MZmine recovers a similar number of analytes as before (Table 1). Surprisingly, the other two software perform worse than their lowest number of analytes found in the above-mentioned data, indicating that not only method sensitivity plays a role, but other factors can also lead to false-negatives (in higher QC samples). Most likely, still low signal-to-noise ratios and/or low abundant peaks for particular compounds, such as, e.g., alanine, can be considered responsible for this phenomenon, also in higher QC levels. In addition, unfavorable peak shapes, e.g. jitter, wider or non-bell-shaped peaks can affect the false-negative rate. Curiously, FullScan data seems to be less affected by this constraint than DDA data, maybe because of the higher number of data points generated per peak. Once again, ProgQi performs the worst, with a considerable gap in its performance also between the two acquisition strategies.

Looking at the total number of features (untargeted), msDial delivers similar numbers for both data acquisition modes. Different trends are observed for MZmine and ProgQi, where DDA yields in far fewer features than FullScan in line with published data [33]. It is also seen that while msDial delivers about 10,000 features, ProgQi finds about half of that in DDA while it finds more than three-fold more features in FullScan data. MZmine on the other hand outperforms msDial and ProgQi massively with two-fold more features in DDA and about five-fold more features in FullScan.

Very slight differences in retained features for complete cases are observed in all the data. Overall, it could be interpreted that ProgQi performs slightly worse than the other two. This is a first indication of how well or also how aggressive all three software perform gap filling. ProgQi offers no parameters to change gap filling, msDial does offer some while MZmine, as with all other steps, offers a huge selection of adjustable parameters. If time and knowledge permits, adjusting those parameters can increase true positive features and hence reduce the data for statistical analysis in a later step. With current technology, it seems unlikely that this leads to a significant decrease in analysis time as computing power is more than sufficient for handling a few ten thousand features simultaneously. If the strategy is to keep as many features in the data as possible, hence increasing the total features by optimizing peak picking parameters towards that goal, we strongly advise having meaningful parameters to filter out some of the more dubious features found. Having as many true positive features as possible seems paramount for multivariate statistical analysis where false positive features would potentially skew statistics in many ways if left uncontrolled.

Considering the total number of identified features, one could conclude a massive advantage of MZmine over the other two. However, the total amount of features in general, is a bad quality indicator for metabolomics data. Given the large number of features detected, manual differentiation between good features ("true positives") and bad features ("false positives") is not feasible for all of them. Nevertheless, such a

differentiation would allow a precise comparison of software performance. We, therefore, performed a sample size estimation, calculating the necessary number of features to manually evaluate the whole dataset representatively. At first glance, the percentage of true features in each dataset shows a stark decrease for FullScan data in general and a very bad performance of ProgQi. These values must be put into perspective by back-calculating the "incidence" with the number of features in each dataset (unfiltered). Considering these absolute numbers, MZmine potentially provides the highest number of true positive features when using FullScan data. msDial does preform exceptionally well in DDA data where it takes first place with fewer initial features and very high true positive rate of 62 %. In FullScan data it can't hold up to the others though with a lower number of features combined with an average 30 % true positive rate. While considering total numbers of true positive features is necessary for perspective, the above-mentioned multivariate statistics or another automated modeling still benefits from higher true positive numbers in one single dataset and hence should profit from high rates found in e.g., msDial DDA data.

As manual comparison is not possible in large datasets anyway, quality parameters that can be applied automatically, e.g., during data filtering, need to be established. Parameters evaluated in the current study, therefore included feature linearity, peak area, and peak width.

The per analyte linearity analysis (not shown) revealed little more than what the previous results already indicated. Interestingly the slope trends for the different software were reproducible considering the data sources. It seems that there is an unexplained gap between FullScan and DDA data in ProgQi analysis while the other software differs from manual integration and from each other but not between acquisition types. With the limited parameters that can be adjusted in a ProgQi analysis, we could not form a definitive hypothesis as to why that gap occurs. While in the other software, things such as m/z isolation width could be considered as potential culprits, this is not the case for ProgQi. The actual issue could be a combination of peak width estimation, peak integration algorithm, and potentially smoothing. While both the ADAP chromatogram builder of MZmine and the peak picking algorithm developed by msDial are rather consistent in both data acquisition types, ProgQi clearly is not.

ProQi sums the abundance of different compound ions (e.g. different adducts attributed to one compound) to the final abundance given in the export. Varying numbers of detected adduct ions in FullScan versus DDA mode might be a valuable explanation for over- or underestimation of a certain compound abundance. Ultimately, we could not confirm this hypothesis though.

Looking at the variability of the software peak integration, MZmine clearly takes the lead. There is minimal variation between QC levels and hence even different MZmine projects. The slight increase in variability in FullScan is consistent with the other software's and might be explained by various factors that also influence the other analyses. One of them being that our FullScan data tends to show more baseline jitter due to the increased number of datapoints in the same timeframe compared to DDA data. In extremer cases this might increase the difficulty for software to correctly interpret peak shape, start and end. Once again ProgQi takes last place in this analysis with some concerning levels of variability, especially in the FullScan dataset. Overall, in line with former works, we would recommend implementing linearity filtering of metabolomics data. A common approach seems to be the introduction of a QC dilution series and the calculation of a correlation between the dilution and feature responses. [24–26] Although the presented data did not allow for such an analysis, this approach would allow to not only filter spiked/targeted analytes but all features of a metabolomics dataset hence providing information on their respective quality and usefulness under given circumstances.

There seems to be little effect of filtering for a minimum area of 300. This was expected as at least in MZmine and msDial minimum peak intensities were defined at reasonable or even low values. The biggest impact of this parameter seems to be for FullScan data in both MZmine and ProgQi. One explanation could be that both datasets contain a huge number of features to begin with, increasing the chance for lower abundant features. Overall, we would recommend thinking about area and peak intensities when filtering but already choosing sensible values for peak picking should limit false positive features regarding low abundancy or incorrectly integrated peaks and hence the effect of this filtering criteria.

Filtering datasets that allow for peak width calculation for sensible peak widths has a major impact on the number of retained data. In our context, a peak width of at least 0.05 min was determined to be sufficient, shorter peak widths are very unlikely to result in reliable features. Looking at the percentage of retained features it is apparent that especially in FullScan data a very high number of features exhibits inadequate peak width. This could either be due to incorrectly integrated peaks or false positive peak picking to begin with. Using this filtering strategy also puts the massive total feature counts observed for MZmine and ProgQi into perspective. Filtering for sensible peak widths could help reduce datasets to sensible features for further statistical analysis. On that note, an implementation of peak width into the export of msDial seems reasonable and needed for future releases.

Lastly, filtering for features that are only present in all samples does make a lot of sense in this experiment since all QC samples should contain the same metabolites. But this also means the effect of this parameter is observably limited.

Considering software performance, MZmine is clearly superior to msDial and ProgQi. It takes less than half of the time in both data sets compared to the other two software with their respective best file format. While we tried to ensure even benchmarking conditions, it is important to stress that msDial and MZmine give options to improve their performance in the parameter's settings while ProgQi does not. Interestingly, both open-source software solutions benefit from a prior file conversion into the open format.mzML, while ProgQi performs better with the original vendor files. Additionally, due to ProgQi licensing options, a physical workstation was needed for the license USB-Stick to be plugged in. In contrast, with the open-source solutions (MZmine and msDial), high-performance cluster infrastructure could be utilized to further speed up the analysis. We are aware that the used data set with eight samples is much smaller than many authentic metabolomics studies. It is well known that processing time does not increase linearly with the addition of more files/larger files, but options like multithreading (available for msDial and MZmine) can improve performance drastically for large batches. Other recommendations to speed up data processing time, might be upgrading available memory (size, then speed) and the use of fast storage (e.g. SSD or even NVMe) as also seen in other studies [50].

5. Conclusion

Overall, our software comparison showed that the performance of msDial and MZmine are close, while ProgQi falls behind current open-source software. Especially considering the steep price point of ProgQi, this is disappointing. From a user-oriented standpoint, msDial comes closer to the sleek-looking and rather straightforwardly organized user interface of ProgQi. We believe the expertise required to successfully generate meaningful and useful data is lowest in msDial and ProgQi while MZmine can be intimidating and hard to learn. Nevertheless, new software releases during our study (e.g., MZmine version 3.0) significantly increased the software performance along with the user interface and workflow. MsDial version 5.0 and future versions of MZmine can likely further improve on certain shortcomings mentioned in this work.

With regard to the quality criteria of metabolome data, we could show again that the total number of features found is not a suitable parameter to judge the peak picking performance of a software. Especially once multivariate statistics are considered, the "garbage in – garbage out" principle holds true. It is, therefore, essential to minimize false positive features in the input data for these analyses. Various

evaluations were made to further understand, improve, and judge datasets generated by msDial, MZmine, and ProgQi. Next to reasonable area filtering, we found that peak width is a very insightful parameter that should be used for data quality in metabolome experiments, although it is not yet implemented in the msDial export. We also recommend including QC samples of the desired matrix, along with QC dilutions to analyze the linearity of all features in a metabolomics dataset. These samples shall be looked at in a targeted fashion (analytes known to be present/spiked etc.). This deepens the understanding of potential peak width, m/z accuracy (also over several samples and concentrations), and potential regions of interest in the chromatography. Using that knowledge, researchers are encouraged to set their choice of software parameters, e.g. r/t tolerance, in line with reasonable values found before. The software's original feature list output should then be filtered for the mentioned quality criteria prior to further (statistical) analysis. Considering msDial and MZmine, our data does not agree with the literature promoting the use of FullScan over DDA to increase meaningful (e.g., true positive) features in metabolomics data. These results, in combination with the time, cost, and material savings when measuring with DDA leave us to believe that initial screenings can readily be performed using DDA.

With our experience from this study, we would like to promote the following standards for benchmarking a metabolomics (data processing) workflow: The inclusion of QC samples with known or even spiked analyte concentrations. Analysis of linearity in known QC samples either spiked or diluted; ideally both. Analysis of peak width and the likeliness of it for the employed LC-method (reporting high likelihood false positives). Manual evaluation of a subset (power analysis) of features and estimating the true positive rate of the workflow. We would also like to discourage the use of total features found as a meaningful parameter of software performance as it promotes the integration of noise and false positive features. Additionally, reporting the ease of use (e.g. GUI, explanation of parameters, documentation of developers) is highly encouraged as we believe it can help developers improve their tools, enabling users to choose a workflow adequate to their skillset.

While our data evaluation largely focused on peak picking, a complete metabolome workflow also relies on data normalization, statistical evaluation, and particularly compound annotation and/or biological interpretation (e.g., through pathway analysis). Taking these important steps into account was outside the scope of the current manuscript but, of course, might influence the preference for particular software, e.g., because of included normalization or identification procedures. In conclusion, the choice of software in a metabolomics workflow is crucial and should be considered before starting the actual data processing.

CRediT authorship contribution statement

Martina Boxler: Investigation. **Yannick Wartmann:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Andrea Steuer:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Thomas Krämer:** Writing – review & editing, Funding acquisition.

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.

Acknowledgements

The authors would like to thank their colleagues Dr. Lana Brockbals and Maja Keller for her support and express their gratitude to Emma Louise Kessler, MD, for her generous legacy she donated to the Institute of Forensic Medicine at the University of Zurich, Switzerland, for research purposes.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jpba.2024.116302.

References

[1] G.J. Patti, O. Yanes, G. Siuzdak, Metabolomics: the apogee of the omics trilogy, Nat. Rev. Mol. Cell Biol. 13 (2012) 263–269.
[2] E.W. Sayers, E.E. Bolton, J.R. Brister, K. Canese, J. Chan, D.C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B.W. Trawick, K.D. Pruitt, S.T. Sherry, Database resources of the national center for biotechnology information, Nucleic Acids Res. 50 (2022) D20–d26.
[3] A. Zhang, H. Sun, G. Yan, P. Wang, X. Wang, Mass spectrometry-based metabolomics: applications to biomarker and metabolic pathway research, Biomed. Chromatogr. 30 (2016) 7–12.
[4] Y. Lu, C. Chen, Metabolomics: bridging chemistry and biology in drug discovery and development, Curr. Pharmacol. Rep. 3 (2017) 16–25.
[5] M.S. Klein, J. Shearer, Metabolomics and type 2 diabetes: translating basic research into clinical application, J. Diabetes Res. 2016 (2016) 3898502.
[6] A.E. Steuer, L. Brockbals, T. Kraemer, Metabolomic strategies in biomarker research-new approach for indirect identification of drug consumption and sample manipulation in clinical and forensic toxicology? Front. Chem. 7 (2019) 319.
[7] S.K. Manier, M.R. Meyer, Current situation of the metabolomics techniques used for the metabolism studies of new psychoactive substances, Ther. Drug Monit. 42 (2020) 93–97.
[8] M. Szeremeta, K. Pietrowska, A. Niemcunowicz-Janica, A. Kretowski, M. Ciborowski, Applications of metabolomics in forensic toxicology and forensic medicine, Int J. Mol. Sci. 22 (2021).
[9] A.E. Steuer, L. Brockbals, T. Kraemer, Untargeted metabolomics approaches to improve casework in clinical and forensic toxicology—"Where are we standing and where are we heading?", WIREs Forensic Sci. (2021).
[10] C. Chen, F.J. Gonzalez, J.R. Idle, LC-MS-based metabolomics in drug metabolism, Drug Metab. Rev. 39 (2007) 581–597.
[11] M. Cuykx, R.M. Rodrigues, K. Laukens, T. Vanhaecke, A. Covaci, In vitro assessment of hepatotoxicity by metabolomics: a review, Arch. Toxicol. 92 (2018) 3007–3029.
[12] R. Ghanbari, S. Sumner, Using Metabolomics to Investigate Biomarkers of Drug Addiction, Trends Mol. Med. 24 (2018) 197–205.
[13] T. Kind, H. Tsugawa, T. Cajka, Y. Ma, Z. Lai, S.S. Mehta, G. Wohlgemuth, D. K. Barupal, M.R. Showalter, M. Arita, O. Fiehn, Identification of small molecules using accurate mass MS/MS search, Mass Spectrom. Rev. 37 (2018) 513–532.
[14] W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D. B. Kell, R. Goodacre, C. Human Serum Metabolome, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, Nat. Protoc. 6 (2011) 1060–1083.
[15] O.C. Zeki, C.C. Eylem, T. Recber, S. Kir, E. Nemutlu, Integration of GC-MS and LC-MS for untargeted metabolomics profiling, J. Pharm. Biomed. Anal. 190 (2020) 113509.
[16] R. Spicer, R.M. Salek, P. Moreno, D. Canueto, C. Steinbeck, Navigating freely-available software tools for metabolomics analysis, Metabolomics 13 (2017) 106.
[17] A.E. Steuer, L. Brockbals, T. Kraemer, Untargeted metabolomics approaches to improve casework in clinical and forensic toxicology—"Where are we standing and where are we heading? WIREs Forensic Sci. 4 (2022) e1449.
[18] S. Hemmer, S.K. Manier, S. Fischmann, F. Westphal,, L. Wagmann, M.R. Meyer, Comparison of three untargeted data processing workflows for evaluating lc-hrms metabolomics data, Metabolites 10 (2020).
[19] J. Guo, T. Huan, Comparison of full-scan, data-dependent, and data-independent acquisition modes in liquid chromatography–mass spectrometry based untargeted metabolomics, Anal. Chem. 92 (2020) 8072–8080.
[20] L. Wang, F.J. Naser, J.L. Spalding, G.J. Patti, A protocol to compare methods for untargeted metabolomics, Methods Mol. Biol. 1862 (2019) 1–15.
[21] E. Defossez, J. Bourquin, S. von Reuss, S. Rasmann, G. Glauser, Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics, Mass Spectrom. Rev. 42 (2023) 131–143.
[22] F. Fenaille, P. Barbier Saint-Hilaire, K. Rousseau, C. Junot, Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: where do we stand? J. Chromatogr. A 1526 (2017) 1–12.
[23] M.I. Boxler, T.D. Schneider, T. Kraemer, A.E. Steuer, Analytical considerations for (un)-targeted metabolomic studies with special focus on forensic applications, Drug Test. Anal. 11 (2019) 678–696.
[24] V. Croixmarie, T. Umbdenstock, O. Cloarec, A. Moreau, J.-M. Pascussi, C. Boursier-Neyret, B. Walther, Integrated Comparison of Drug-Related and drug-induced ultra performance liquid chromatography/mass spectrometry metabonomic profiles using human hepatocyte cultures, Anal. Chem. 81 (2009) 6061–6069.
[25] M.R. Lewis, J.T.M. Pearce, K. Spagou, M. Green, A.C. Dona, A.H.Y. Yuen, M. David, D.J. Berry, K. Chappell, V. Horneffer-van der Sluis, R. Shaw, S. Lovestone, P. Elliott, J. Shockcor, J.C. Lindon, O. Cloarec, Z. Takats, E. Holmes, J.K. Nicholson, Development and application of ultra-performance liquid chromatography-tof ms

for precision large scale urinary metabolic phenotyping, Anal. Chem. 88 (2016) 9004–9013.
[26] D. Broadhurst, R. Goodacre, S.N. Reinke, J. Kuligowski, I.D. Wilson, M.R. Lewis, W. B. Dunn, Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies, Metabolomics 14 (2018) 72.
[27] Z. Li, Y. Lu, Y. Guo, H. Cao, Q. Wang, W. Shui, Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection, Anal. Chim. Acta 1029 (2018) 50–57.
[28] J. Liao, Y. Zhang, W. Zhang, Y. Zeng, J. Zhao, J. Zhang, T. Yao, H. Li, X. Shen, G. Wu, W. Zhang, Different software processing affects the peak picking and metabolic pathway recognition of metabolomics data, J. Chromatogr. A 1687 (2023) 463700.
[29] O.D. Myers, S.J. Sumner, S. Li, S. Barnes, X. Du, 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. 89 (2017) 8696–8703.
[30] L. Hao, J. Wang, D. Page, S. Asthana, H. Zetterberg, C. Carlsson, O.C. Okonkwo, L. Li, Comparative evaluation of MS-based metabolomics software and its application to preclinical Alzheimer's disease, Sci. Rep. 8 (2018) 9291.
[31] A. Cambiaghi, M. Ferrario, M. Masseroli, Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration, Brief. Bioinform. 18 (2017) 498–510.
[32] A. Marco-Ramell, M. Palau-Rodriguez, A. Alay, S. Tulipani, M. Urpi-Sarda, A. Sanchez-Pla, C. Andres-Lacueva, Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data, BMC Bioinforma. 19 (2018) 1.
[33] J. Guo, T. Huan, Comparison of full-scan, data-dependent, and data-independent acquisition modes in liquid chromatography-mass spectrometry based untargeted metabolomics, Anal. Chem. 92 (2020) 8072–8080.
[34] A. Rafiei, L. Sleno, Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis, Rapid Commun. Mass Spectrom. 29 (2015) 119–127.
[35] O.D. Myers, S.J. Sumner, S. Li, S. Barnes, X. Du, 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. 89 (2017) 8689–8695.