Benchmarking feature quality assurance strategies for non-targeted metabolomics

Yasin El Abiead[1,*], Maximilian Milford[1], Harald Schoeny[1], Mate Rusz[1,2], Reza M Salek[3], and Gunda Koellensperger[1,*]

1 Department of Analytical Chemistry, University of Vienna, Vienna 1090, Austria, 2 Department of Inorganic Chemistry, University of Vienna, Vienna 1090, Austria, and 3 International Agency for Research on Cancer, Section of Nutrition and Metabolism, Lyon 96008, France

ABSTRACT: Automated data pre-processing (DPP) forms the basis of any liquid chromatography-high resolution mass spectrometry-driven non-targeted metabolomics experiment. However, current strategies for quality control of this important step have rarely been investigated or even discussed. We exemplified how reliable benchmark peak lists could be generated for eleven publicly available datasets acquired across different instrumental platforms. Moreover, we demonstrated how these benchmarks can be utilized to derive performance metrics for DPP and tested whether these metrics can be generalized for entire datasets. Relying on this principle, we cross-validated different strategies for quality assurance of DPP, including manual parameter adjustment, variance of replicate injection-based metrics, unsupervised clustering performance, automated parameter optimization, and deep learning-based classification of chromatographic peaks. Overall, we want to highlight the importance of assessing DPP performance on a regular basis.

Liquid chromatography – high-resolution mass spectrometry (LC-HRMS) is the most widely used tool for advancing research in non-targeted metabolomics, lipidomics, and exposomics. A crucial step in hypothesis-free extraction of information from complex LC-HRMS data is automated non-targeted data pre-processing (NPP). NPP typically consists of a cascade of steps, including chromatogram extraction, chromatographic peak picking, peak alignment, and aligned feature processing (e.g., gap-filling) leading to so-called feature-tables (FT; Figure S1), which typically form the basis of any subsequent interpretation. Over the past years, numerous algorithms and software packages enabling NPP have been published (e.g., XCMS[1], XCMS-online[2], MZmine 2[3], MS-DIAL[4], El-MAVEN[5], OpenMS[6], and others[7,8]). Reproducing results on different tools and optimizing parameters has been recognized as one of the main challenges in non-targeted data processing, impacting data interpretation and ultimately study outcome.[9–15] The emerging unease regarding the under-utilization of data[12,13], calls for evaluation strategies that enable benchmarking of different tools, algorithms, and parameter choices on a regular basis. Undeniably, validation of LC-HRMS NPP poses several challenges. First, the true number of peaks (and their properties such as area, height, etc.) in a raw dataset is typically unknown. This is even true when measuring a defined set of LC-HRMS grade analytical standards due to unavoidable chemical impurities, contaminations, unexpected adducts, and gradient artifacts. Second, no accepted gold-standard NPP method which would allow the generation of reliable reference FT as benchmarks exists. Finally, datasets differ in their characteristics (peak widths, mass precision, etc.), which implies that conclusions on NPP performance drawn from a benchmark might only allow conclusions for other datasets with similar characteristics. State-of-the-art NPP reliability assessment strategies include peak/feature classification[16–18], unsupervised clustering[18–21], and benchmark recovery-based[9,22,23] techniques, with the first two being applied almost routinely, while recovery-based approaches were mainly limited to dedicated NPP assessment studies. Up to date, there is no standardized catalog of metrics enabling the assessment and reporting of general figures of merit. In fact, the opposite is the case: there is vast heterogeneity in used metrics and their ability to describe NPP performance.

Figure 1 provides an overview of current classification-based NPP assessment strategies, their metrics, and general capabilities. Generally, classification discerns reliable from unreliable peaks/features (e.g., reliable LC-HRMS peaks should co-elute with isotopologues identifiable via specific mz distances), and many principles can be easily implemented as NPP performance tests. However, key aspects of NPP quality remain blind spots. First, for peaks/features classified as unreliable, it is generally not possible to differentiate between cases where chromatographic noise has been erroneously picked as peak and cases where chromatographic peaks have been picked poorly. Moreover, the majority of principles allowing to assess the quality of reported abundances yields a checkpoint only after alignment. Therefore, it is difficult to trace the problems back to the peak picking or alignment step. Finally, classification-based methods do not assess the proportion of undetected peaks (false negatives).

NPP assessment through comparison with targeted data evaluation, denoted as benchmark recovery-based approaches, has the potential to investigate these blind spots. While the concept is straightforward, it is rarely applied in routine non-targeted experiments, as its implementation can be tedious. This is because benchmark generation requires meticulous manual curation of peaks, which can be very time-consuming and is often considered to be too subjective for ground-truth generation. In fact, a recent study showed that three experts in mass spectrometry strongly disagreed on what constitutes an actual chromatographic peak in ~20% of cases (n = 1071)[24], demonstrating how vague boundaries between differences in opinion carry the risk of overinterpreting differences between benchmarks and NPP results. Moreover, the manual work of benchmark curation leads to rather small sets of benchmark peak lists, potentially hampering the representativeness of the benchmarks for the whole dataset.[14,21]

We recently introduced mzRAPP, a tool that allows simplifying benchmark recovery-based NPP assessment while also increasing its reliability and providing functionality for their

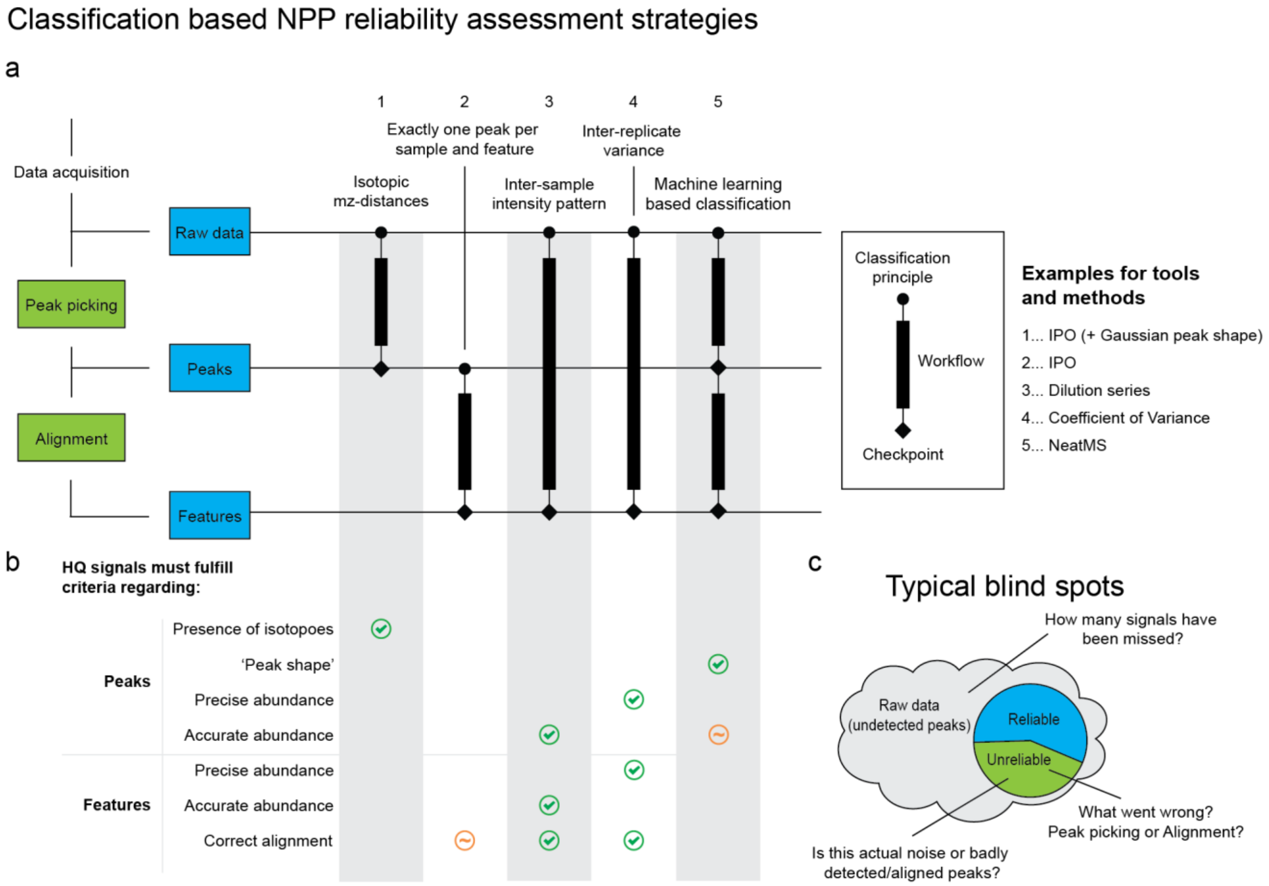
Classification based NPP reliability assessment strategies
a
1
2
3
4
5
Exactly one peak per sample and feature
Inter-replicate variance
Data acquisition
Isotopic mz-distances
Inter-sample intensity pattern
Machine learning based classification
Raw data
Peak picking
Peaks
Alignment
Features
Classification principle
Workflow
Checkpoint
Examples for tools and methods
1... IPO (+ Gaussian peak shape)
2... IPO
3... Dilution series
4... Coefficient of Variance
5... NeatMS
b
HQ signals must fulfill criteria regarding:
Peaks
Presence of isotopoes
'Peak shape'
Precise abundance
Accurate abundance
Features
Precise abundance
Accurate abundance
Correct alignment
c
Typical blind spots
How many signals have been missed?
Raw data (undetected peaks)
Reliable
Unreliable
What went wrong? Peak picking or Alignment?
Is this actual noise or badly detected/aligned peaks?

Figure 1. Selected strategies for classification-based performance assessment for non-targeted data pre-processing (NPP) of LC-HRMS data. (a) Five principles distinguishing reliable high quality (HQ) from unreliable peaks/features have been summarized. While some approaches require only peak picking for their application, others can only be utilized after alignment. (b) Different principles allow for the assessment of data properties. Notably, in cases where multiple criteria have to be fulfilled, all of them have to pass for a peak/feature to be considered HQ. (c) Typically, classification-based approaches suffer from a number of blind spots.

utilization.[25] Briefly, it takes the outputs of traditional targeted metabolomics data evaluation (molecular formulas with retention time boundaries per sample) as an input and automatically increases the number and reliability of benchmark peaks via consideration of isotopologues. In this way, benchmark peak abundances can be confirmed by predicted isotopologue ratios (IRs), rather than solely by human judgment. Finally, the consideration of isotopic relations allows the introduction of performance metrics dissecting the individual steps of NPP (e.g., see Figure S2).

NPP results obtained following the steps of peak-picking, alignment, and gap-filling, respectively are benchmarked using metrics such as the recovery of benchmark peaks and isotope ratio biases (IRbs). The latter reflects the accuracy of the peak abundances obtained as visualized in Figure Figure S2a. As another key advantage, metrics on the NPP alignment process can be retrieved without requiring tedious curation of the correct alignment of all benchmark peaks. By simply counting cases where the alignment performed on a given isotopologue is in conflict with the alignment of another isotopologue of the same molecular species (see Figure S2b). Building upon these principles, mzRAPP generates benchmarks and performs reliability NPP assessment at different stages of the process (picked peaks, peak alignment, and processed features). Currently, mzRAPP supports outputs from more than five of the most commonly used NPP tools (XCMS, XCMS-online, MZmine 2, MS-DIAL, OpenMS, El-MAVEN, MetaboanalystR 3.0[26], etc.).

In this work, we show the power and necessity of this novel validation scheme. We scrutinize current NPP execution and assessment strategies by adding a new validation layer as enabled by representative high-quality benchmarks created case by case. Moreover, we explore the most common NPP optimization/assessment strategies, including the most recent NPP classification approaches based on artificial intelligence, by showing their capability and limitations in an unprecedented manner. For the first time, cross-validation, including the quantitative dimension of features/peaks, is within reach in the otherwise black-box-like environment of NPP.

Methods

Datasets. Datasets used for benchmark generation were downloaded from Metabolights[27] or thankfully provided by the authors of the respective studies.[9,22,24,28–31] References and/or repository IDs are provided in column ‘Reference’ of Table S1. Where not already provided as centroided mzML files, raw files were centroided and converted to mzML format via ProteoWizards msConvert[32] (version 3.0.21045-7732b6429). Names of raw files used are provided in Table S2. Files acquired via fast polarity switching were filtered to contain only positive scans.

Benchmark generation. Targeted data evaluation was performed via Skyline[33] (version 21.1.0.146) and for the most abundant isotopologue of each molecule, for which retention time and molecular formula were known, in all datasets. Then, manual set retention time boundaries were exported for each molecule and mzML file. These and the mzML files formed the input for mzRAPP, which also extracted other predictable isotopologues for each molecular formula. Only isotopologues with a Pearson Correlation Coefficient (PCC) > 0.85, below an isotopologue ratio bias (as calculated by peak areas) of 35%, an isotopologue ratio bias (as calculated by peak heights) of 30%, and a difference between ratio bias (height) and ratio bias (area) below 30% points were accepted.

Extraction of non-targeted data pre-processing performance metrics. Extraction of NPP performance metrics was conducted via mzRAPP (version 1.1.6). Exact criteria and rules for matching between signals of the benchmark and those of the unaligned and aligned NPP outputs can be found in the original mzRAPP publication[25] and on Github (https://github.com/YasinEl/mzRAPP). Isotopologue ratio biases (IRbs), as calculated from NPP outputs, were considered to be recovered if they were less than 20 %points higher than respective BM IRbs. Confidence intervals (confidence level = 0.99), for all NPP metrics were derived via bootstrapping of benchmark molecules (R = 1000) using the boot package (version 1.3-28).

Application of non-targeted data pre-processing. All data pre-processing experiments were performed via XCMS3 (version 3.14.1) using R 4.1.0 and MZmine 2 (version 2.53). Parameter optimizations were performed manually or via automated optimization tools. The automated optimization tools were IPO[17] (version 1.18.0), AutoTuner[34] (version 1.6.0), MetaboanalystR 3.0[26], and SLAW[35] (version 1.0.0). Classification of peaks of peaks by quality was performed via NeatMS version (0.9), which was run via Python 3.7. For the parameter sensitivity study, the coefficient of variance investigation, and the parameter optimization dataset (DS) 5 was processed. For the unsupervised clustering investigation, and the assessment of NeatMS DS 1 was processed. Additional details on set parameters for all performed studies are given in the supplementary material.

Data analysis and figures. All further data analysis was performed using R (version 4.1.0) and R studio (version 1.4.1717) using the data.table package. Plots were generated using ggplot2, patchwork, and ggradar. Figures and diagrams were further processed using Adobe Illustrator (25.3.1).

Results and Discussion

The quality of automated benchmark curation and extension. Ideally, a BM used for NPP assessment should be produced from the same dataset or a dataset generated via the same instrumental platform as the dataset of interest. In this study, BMs from 11 different public and in-house raw datasets (listed in Table S1) were generated via mzRAPP. The case by case generated BMs covered five different MS-systems, coupled to hydrophilic interaction chromatography (HILIC) or reversed-phase chromatography (RPC), as well as different sample types (including analytical standard mixtures, blood serum, red blood cell extracts, and cell culture extracts) and compound classes (polar metabolites, lipids, and exogenous small molecules). Targeted extraction of the most abundant isotopologue of each known molecule was done manually, but was automatically extended to all lower-abundance isotopologues. Quantitative properties of the thereby generated benchmarks are visualized in Figure S3. In Figure S3a all 46742 BM peak areas of low abundant isotopologues (LAITs) were plotted against the area predicted from the respective most abundant isotopologue (MAIT), showing excellent linearity. It should be noted that only peaks within the linear dynamic range of the instrumental platform were added to the BMs. Figure S3b shows the absolute peak area bias of all LAITs, with 94% of all calculated isotopologue ratio biases < 25%. Comparison of biases (Figure S3c) as calculated via peak areas versus peak heights (which are generally more robust as they are not affected by poor setting of RT boundaries) revealed good agreement, further strengthening the evidence of accurate "ground-truth" for an extensive number of peaks. Finally, for DS1, the benchmark reliability was evaluated upon comparison with reported fold changes assessed in an independent laboratory. Figure S3d shows the excellent reproducibility for the mzRAPP approach applied here. For this specific dataset, the number of peaks with reliable quantitative properties increased by > 200% by integrating LAITs. The addition of LAITs increased not only the BM size but also the covered dynamic range. Figure S4a and b quantifies this significant extension for all eleven investigated datasets.

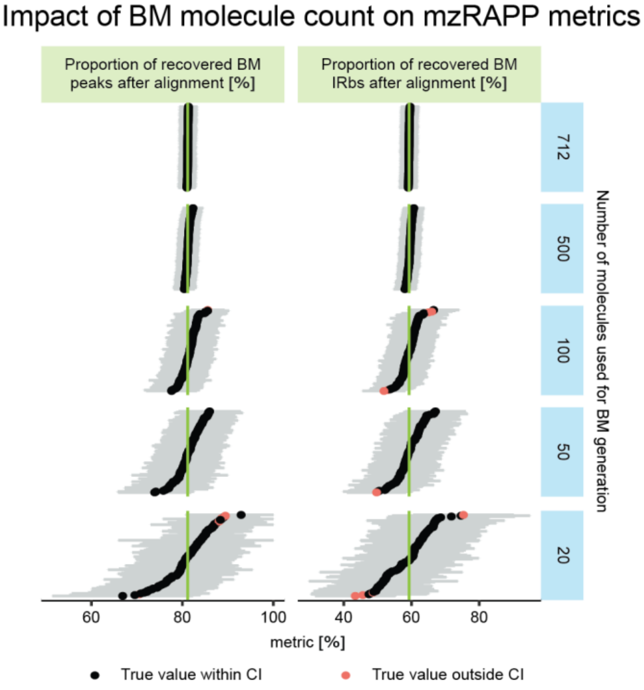

Figure 2. A benchmark was generated for dataset 1 considering 712 molecules and utilized for NPP assessment. Stepwise reduction of molecules used for BM generation (molecules were sampled at random from all 712 molecules (n = 100) for molecule numbers ranging from 20 to 712) largely affected the size of confidence intervals (CI; as estimated by mzRAPP via bootstrapping with R = 1000 and confidence level = 0.99), while true metrics, as estimated from the original BM (green lines) stayed largely within within the reported CI of respective NPP assessments. (CI values above 100 were reduced to 100)

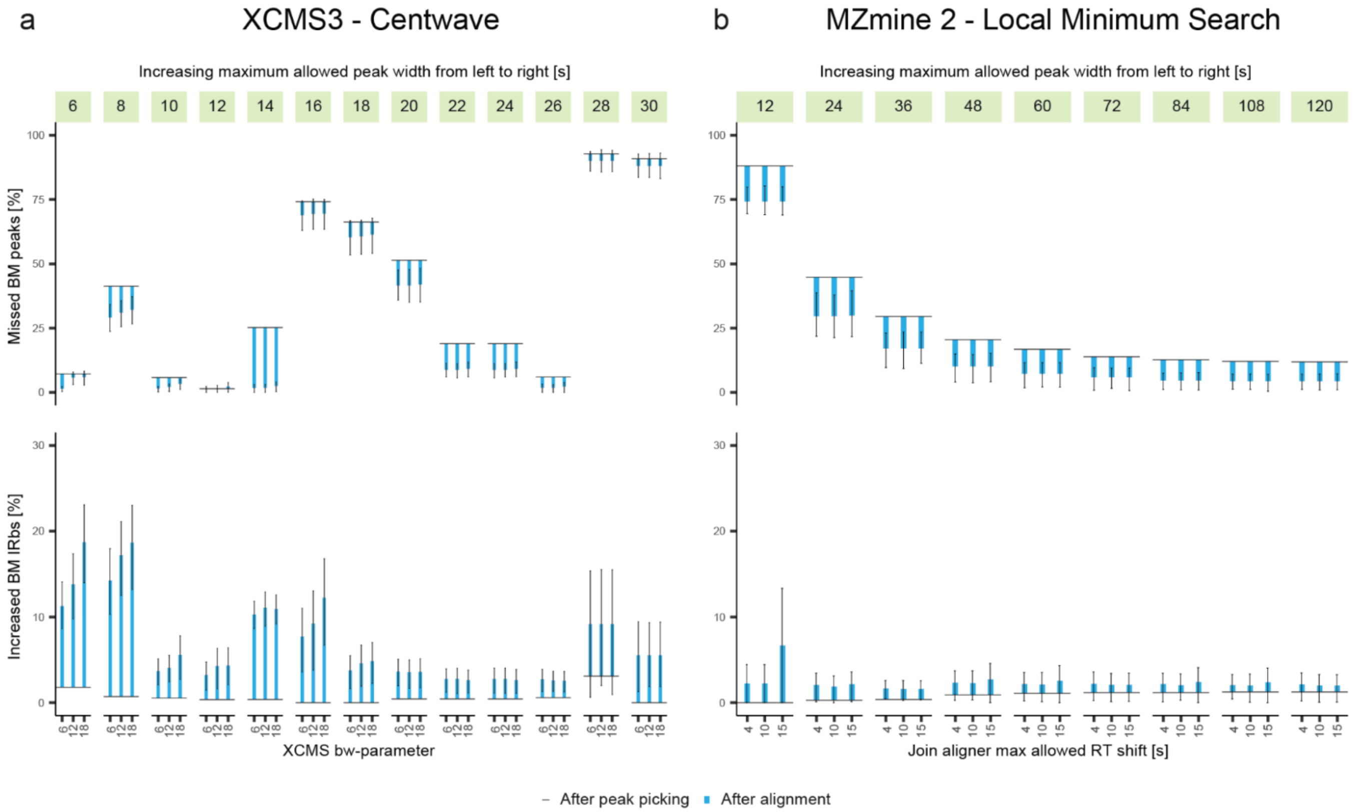

Figure 3. Multiple non-targeted data pre-processing experiments with parameter sweeps for the peak-picking and the alignment step were performed on the same dataset via XCMS3 and MZmine 2. The peak picking parameter referring to the maximum peak width allowed (MPW) was incrementally increased for both tools, with three retention time tolerances tested for each peak picking step. (a) The proportion of recovered BM peaks was strongly and often abruptly affected by the MPW parameter. While the proportion of increased isotopologue ratio biases (IRbs) was more strongly influenced after alignment, this effect was primarily dependent on the MPW parameter during peak picking. (b) In MZmine 2, an increase of the MPW never led to a reduction in the proportion of recovered BM peaks.

Peak metrics such as the full width half maximum (FWHM) of chromatographic peaks and the mass precision given by mz ranges of individual peaks are important for any dataset to be analyzed via NPP. In fact, most NPP tools require parameters corresponding to these variables to be set for any dataset to be processed. Therefore, it is worth noting that generated BMs showed large differences in all these metrics as a result of different measurement methods (see Figure S4c and d). Next to these characteristics, the peak shape as reflected by the zigzag index[39], sharpness[40], and other metrics (Figure S5) show significant differences across investigated datasets. This highlights the importance of using benchmarks very similar to the dataset of interest (and at best generated from the very same dataset) for potential NPP assessment studies.

Ultimately, benchmarks can be utilized to derive performance metrics for NPP experiments conducted on the same raw datasets. As outlined above, mzRAPP enables automated assessment of the proportion of recovered benchmark peaks and the proportion of recovered isotopologue ratio biases (IRbs) before and after alignment as key NPP performance metrics. Moreover, other metrics such as the proportion of NPP peaks with reported integration boundaries close to the intensity maximum of a benchmark peak are reported.

Finally, the metrics derived from the benchmark should be translatable into an actual proportion for the underlying dataset. Representative sampling of the benchmark peaks/features is a prerequisite. To show the validity of our approach, we estimated confidence intervals (CI) for all metrics by bootstrapping benchmark molecules. In Figure 2 and Figure S6 we show how the number of molecules used for BM generation affects metrics derived from NPP reliability assessment. This was done by bootstrapping different numbers of molecules from BM 1 (containing 712 molecules and > 30000 peaks). It can be observed how a reduction of the number of BM molecules increases the CI while the assessed metric was in agreement in almost all cases. Therefore, even < 100 BM molecules can lead to reasonable estimates of the performance of NPP, as long as the increase in CI can be accepted.

Sensitivity of NPP parameters. NPP requires the adaptation of different parameters to the analyzed dataset. These parameters can appear more or less intuitive to users with different scientific backgrounds and experiences. Generally, parameters involving expected chromatographic peak widths and retention time shifts are often considered to be among the more intuitive parameters. In the following, we showcase examples that emphasize the need for case-by-case benchmarking strategies, as even intuitive parameter settings in the NPP could have an adverse impact on the NPP output.

Figure 3a shows how a stepwise increase of XCMS's centwave's maximum peak width parameter (MPW) using 2 s increments heavily affected the proportion of missed BM peaks and increased BM isotopologue ratio biases (IRbs). In the most extreme case, an increase in MPW from 26 to 28 s led to an increase in the proportion of missed BM peaks (before alignment) from 6 to 93%. Considering that the median of BM peaks full-width half maxima (FWHM) ranged from 4 to 18 s with a median of ~7 s, there was no trivial dependence of the optimal MPW on the FWHM of peaks to be detected. While fewer peaks were missed after alignment and gap-filling, this improvement was insufficient to make up for errors introduced during peak detection. It is worth noting that even in cases where gap filling recovered most peaks, such as with an MPW of 14 s, the resulting peak areas led to worse IRbs than when peaks were already detected in the peak detection step (e.g., with MPW set to 12 s). While the highest observed retention time shift in the BM peaks was below 10 s, the maximum allowed shift, as set via the bandwidth (bw) parameter in the group density algorithm, did not affect NPP results to the same extent as MPW. Interestingly, there was almost no effect of the set MPW on IRbs after peak picking. However, there was a significant impact on IRbs after peak alignment and gap-filling, which depended on the MPW set during the peak picking step rather than set alignment parameters.

The same dataset was processed via MZmine 2 and incremental increase of the MPW using its Local Minimum Search peak detection algorithm as visualized in Figure 3b. Here, the relationship between the set MPW and the proportion of recovered peaks was found to be as expected, with more BM peaks being recovered with increasing MPW. However, the absolute MPW necessary to reach the optimum of ~20% missing BM peaks after peak detection was > 72 s, a factor of ~10 higher than the median BM FWHM of ~7 s. While this algorithm assessed the MPW at the base of peaks rather than at half maximum, this might still lead to problems since the base width of a chromatographic peak can vary widely for real-world applications. The IRbs were almost unaffected by the MPW, with only one outlier for an MPW of 12 s and a retention time tolerance of 15 s.

For this specific data set, the optimum of all parameter sets tested was found by XCMS (MPW of 12 s and retention time tolerance of 6 s, leading to a proportion of missed BM peaks < 1% and IRbs < 5%). It should be noted that this finding cannot be generalized, but it holds true for the processed data set, representing a use case of parameter adjustment. In fact, the conclusion that MZmine 2 is generally underperforming as compared to XCMS would be wrong since testing the entire parameter space was beyond the scope of this study. The example clarifies that the common practice of manual parameter adjustment to metrics derived from the analytical performance of the instrumental setup can lead to suboptimal NPP results.

Application of NPP parameter optimization tools. Current NPP parameter optimization tools (as implemented in IPO[17], AutoTuner[34], MetaboanalystR 3.0[26], and SLAW[35]) undoubtedly facilitate the NPP step in non-targeted analysis and improve quality. Here we test the classification-based tools applying our benchmark-recovery approach. This way, the otherwise missing metrics of missing peaks and accuracy of peak abundances are validated. Figure 4 compares the quality of NPP-parameter optimization performed for the NPP of DS 5 via different metrics as exported by mzRAPP. As can be seen, the differences between the optimization attempts were observable for the proportion of missed benchmark peaks and increased IRbs before and after alignment, as well as the proportion of BM peaks leading to split peaks. It turned out that the initially defined values for the parameter optimization process is crucial and was unique for each tool. For example, in the case of IPO, manual adjustment of starting parameters led to a decrease in the proportion of BM peaks missing after alignment from ~25% to ~1%. Again, this test emphasizes that NPP evaluation is not redundant when using automated parameter optimization algorithms.

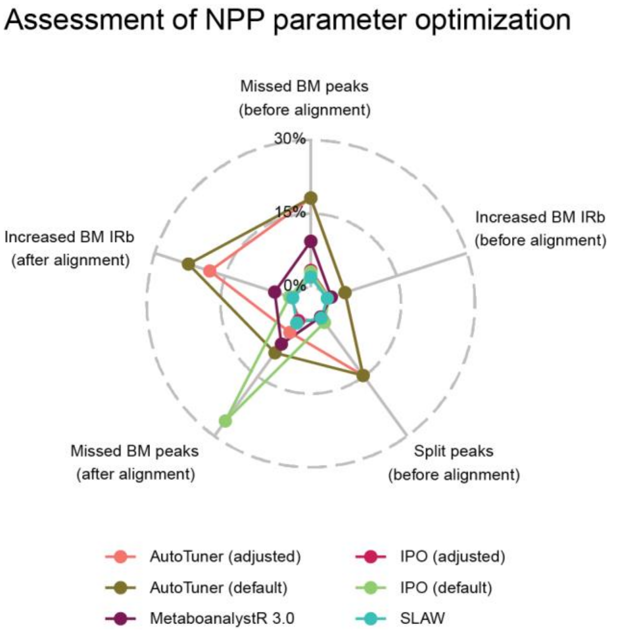

Figure 4. Non-targeted data pre-processing (NPP) parameters for the processing of dataset 5 have been optimized using different optimization tools (IPO and AutoTuner, both adjusted and default, as well as MetaboanalystR 3.0 and SLAW). Outcomes were assessed via five benchmark recovery-based metrics, namely the proportion of missed benchmark (BM) peaks (before and after alignment), the proportion of increased isotopologue ratio biases (before and after alignment), and the proportion of split peaks (before alignment). Confidence intervals of all metrics for the underlying dataset are given in Table S3.

NPP assessment or filtering via Coefficient of Variance. A widely accepted metric in NPP assessment is the Coefficient of Variance (CV) obtained from replicate measurements. It is common practice to report the number or proportion of NPP features with a Coefficient of Variance (CV) below a certain threshold (commonly 30%). In Figure 5, this otherwise straightforward approach is investigated. CV-based quality metrics are plotted versus proportions of recovered BM peaks and IRbs for a total of 78 NPP experiments (using different values for the maximum allowed peak width in the peak-picking step and different values for bw in the peak alignment step) performed on DS 5. The number and proportion of features with a CV < 30% (nCV30 and pCV30, respectively) were assessed. For a given bw setting, the metric nCV30 was well correlated with the proportion of BM peaks recovered post-alignment (e.g., PCC = 0.9

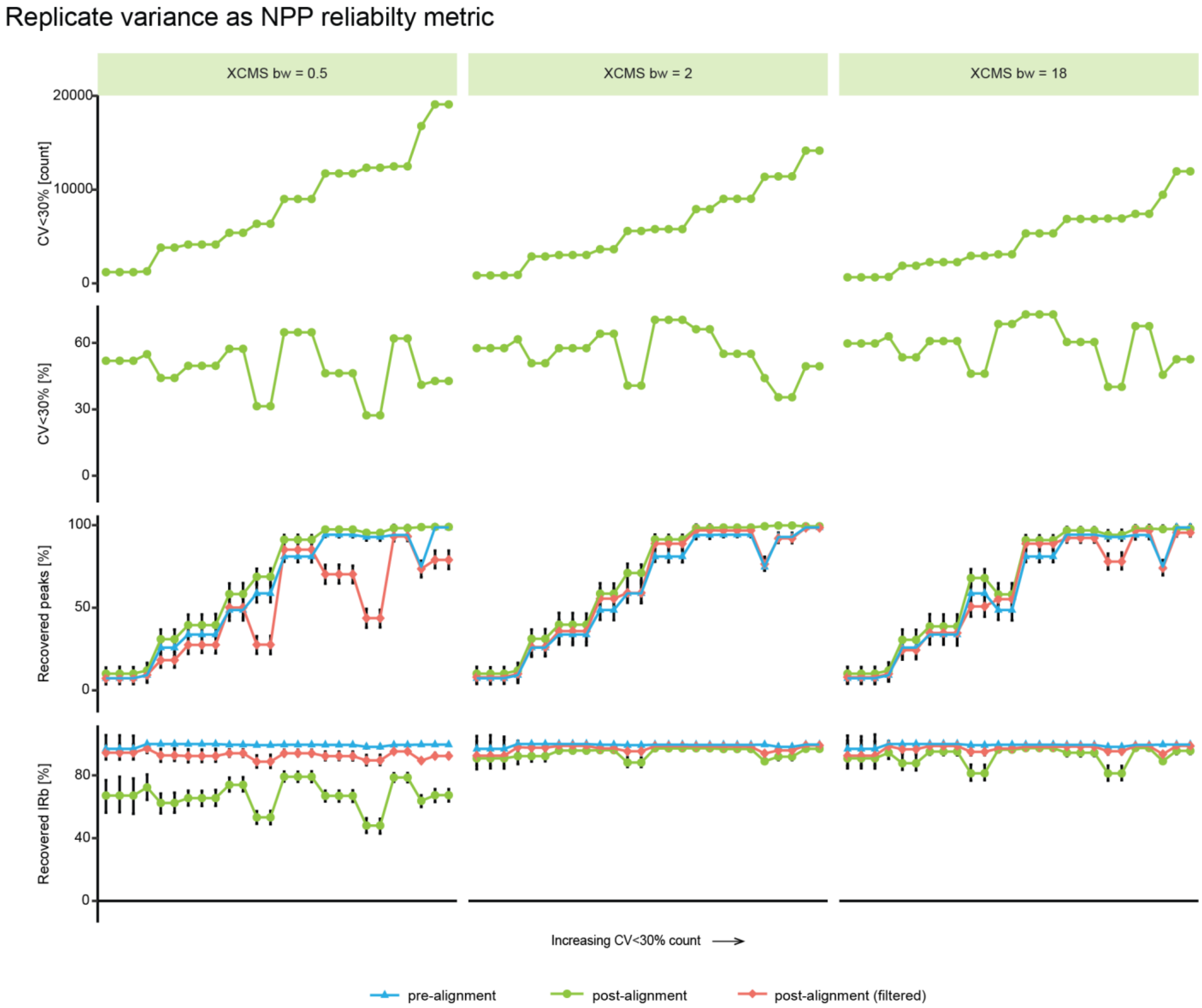

Figure 5. A dataset consisting of 9 replicate injections was processed via XCMS using different values for the maximum allowed peak width (MPW) parameter and the bandwidth (bw) parameter, leading to a total of 78 non-targeted data pre-processing experiments. Four different quality metrics including the number of features with CV < 30%, the proportion of features with CV < 30%, and the proportion of recovered benchmark peaks and isotopologue ratio biases (IRb) (pre-alignment, post-alignment, and post-filtering (only features without missing values and with CV < 30%)) were then plotted (sorted by increasing number of features with CV < 30%).

for bw = 0.5), while the proportion of features with CV < 30% reflected the proportion of IRb recovered post-alignment (e.g., PCC = 0.97 for bw = 0.5). Filtering NPP features to remove features with missing peaks and setting a threshold of CV > 30% led to an increase in the proportion of recovered IRbs (post-alignment) in all cases. In contrast, the proportion of overall recovered peaks decreased. Thus, the trends in nCV30 were indicative of the proportion of found peaks, while the trends in pCV30 corresponded to the proportion of high-quality peaks. The application of this filter successfully removed low-quality features contributing to increased IRbs. This highlights how classification-based NPP assessment approaches do not distinguish between noise and badly detected peaks, as discussed above (as well as in Figure 1). It should be noted that the high correlation of pCV30 with the proportion of recovered IRb (post-alignment) persisted only as long as the bw parameter was kept constant. For example, while the trends of pCV30 were very similar in all settings of bw, the proportion of recovered IRb (post-alignment) for bw = 0.5 decreased for all MPW values. As a consequence, pCV30 is not a suitable metric to evaluate whether one given NPP result has higher quality than another. The same is true for nCV30. In fact, the highest nCV30 (for bw = 0.5) would have led to ~75% recovered IRbs (after alignment), which is far from the optimum of approximately 100% (with values > 2 or 18). Thus, both CV metrics have their valid role in quality assessment. However, simple NPP optimization based on CV metrics is precluded. Neither optimizing nCV30 nor the pCV30 led to a complete recovery of BM peaks and BM IRb and thus to an optimum result.

NPP assessment via unsupervised clustering. Unsupervised clustering approaches such as Principal Component Analysis (PCA) are an integral part of non-targeted workflows and their role in discovery is undisputed. However, they are not suitable for NPP-quality assessment, as emphasized in Figure S7 and Figure S8. PCAs performed on the NPP outputs showed good separation of sample groups, not reflecting the benchmark recovery metrics, and thus not validating the reliability of the NPP output.

Application of peak classification via artificial intelligence. Novel tools such as NeatMS use artificial intelligence for the classification of NPP peaks by their quality. As a major breakthrough, noise removal is accomplished without relying on replicate injections. Successful application of machine learning algorithms requires good training data, which (in the case of NeatMS) have to be labeled by users with different skill sets. In this work, we scrutinized NeatMS. NPP peaks generated from nine NPP experiments performed on the same dataset (DS 1; containing 10 samples), were classified accordingly, into three categories 'High quality', 'Low quality' and 'Noise'. We then applied different filters to the aligned NPP features and required them to contain 0, 1, 3, 5, 8, or 10 'High quality' peaks. Subsequently, the proportion of recovered peaks and IRbs after alignment was assessed. For this purpose, we filtered our BM to contain only features with peaks in all 10 samples. As can be seen in Figure 6, removing all NPP features which did not include at least 1 'High quality' peak reduced the number of NPP features by ~40 to ~60%, while having almost no effect on the proportion of recovered BM peaks or IRbs, demonstrating how NeatMS can be applied successfully for removing false positives from NPP results. However, requiring more 'High quality' peaks reduced the proportion of recovered BM peaks by multiple %points in many cases. When all 10 samples were required to contain only 'High quality' peaks for a feature to be retained, the proportion of recovered BM peaks dropped <10% in all cases. Our validation confirms that tools such as NeatMS for efficiently removing false-positive signals from NPP results significantly advance NPP. Despite this undisputed role, the quality and size of training data strongly affects the procedure and is defined by a user, case by case. Thus, independent validation continues to be of great value.

Conclusion

The study emphasized the need of case-by-case optimization and validation of NPP. Due to the described obstacles in manually adjusting parameters, even when applying cutting-edge optimization tools, we reason that NPP performance should be evaluated on a regular basis. Moreover, discussed problems in finding NPP optima via strategies such as unsupervised clustering or Coefficient of Variance-based metrics entail the necessity for alternative assessment methods. NPP can perform in unpredictable ways, and outputs should be assessed against a solid ground-truth which cannot be generalized in the form of a golden dataset, against which all NPP algorithms are optimized. We demonstrated how case-by-case benchmark recovery-based approaches can satisfy this need. The fact that benchmark information can be based on a solid foundation of orthogonal information allows for increased reliability and accuracy. mzRAPP can produce representative and reliable benchmarks for each investigated dataset, which can be integrated for routine validation of non-targeted NPP. Finally, the performance ranking of different NPP tools (as reported elsewhere) is not useful.

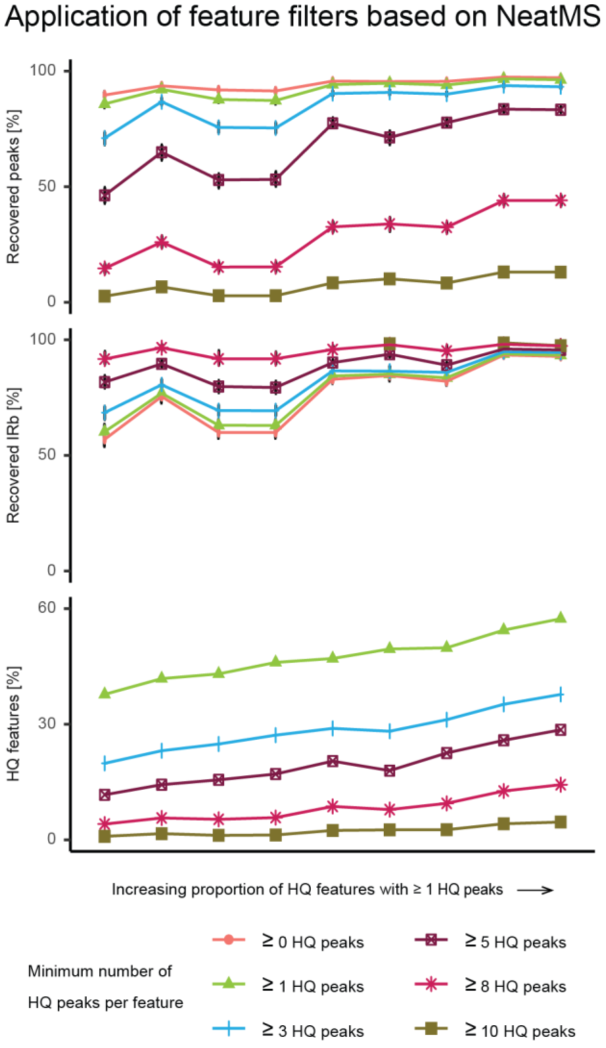

Figure 6. Dataset 1, consistent of 10 samples, was processed with 9 different sets of XCMS-parameters. All peaks produced via XCMS were classified by NeatMS into different categories, including high quality (HQ), or noise. Different numbers of peaks within an aligned feature were required to be of HQ for a feature to be retained. The plot on the bottom shows the proportion of all XCMS features satisfying these criteria for each parameter set. The plots above show metrics on the proportions of recovered benchmark peaks and recovered isotopologue ratio biases (IRbs). The x axis was sorted by increasing values of HQ features [%] for more than or equal to 1 HQ peaks.

ASSOCIATED CONTENT

Supporting Information

Supplemental Methods, Figures S1 – S8, and Table S1 (PDF)
Table S2 (CSV)
Table S3 (CSV)

AUTHOR INFORMATION

Corresponding Authors

Yasin El Abiead – Department of Analytical Chemistry, University of Vienna, 1090 Vienna, Austria;
https://orcid.org/0000-0003-4392-7706;
Email: yasin.el.abiead@univie.ac.at
Gunda Koellensperger – Department of Analytical Chemistry, University of Vienna, 1090 Vienna, Austria;
https://orcid.org/0000-0002-1460-4919;
Email: gunda.koellensperger@univie.ac.at

Author Contributions

The manuscript was written through contributions of all authors. / All authors have given approval to the final version of the manuscript.

Notes

The authors declare no conflict of interest. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization. R.S. would like to acknowledge support from ANR-19-CE45-0021 (MetClassNet) grant.

ACKNOWLEDGMENT

The authors would like to acknowledge Yoann Gloaguen for his advice regarding the application of NeatMS.

REFERENCES

(1) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78* (3), 779–787. https://doi.org/10.1021/ac051437y.

(2) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal. Chem.* **2012**, *84* (11), 5035–5039. https://doi.org/10.1021/ac300698c.

(3) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. *BMC Bioinformatics* **2010**, *11* (1), 395. https://doi.org/10.1186/1471-2105-11-395.

(4) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis. *Nat. Methods* **2015**, *12* (6), 523–526. https://doi.org/10.1038/nmeth.3393.

(5) Agrawal, S.; Kumar, S.; Sehgal, R.; George, S.; Gupta, R.; Poddar, S.; Jha, A.; Pathak, S. El-MAVEN: A Fast, Robust, and User-Friendly Mass Spectrometry Data Processing Engine for Metabolomics. In *High-Throughput Metabolomics: Methods and Protocols*; D'Alessandro, A., Ed.; Methods in Molecular Biology; Springer: New York, NY, 2019; pp 301–321. https://doi.org/10.1007/978-1-4939-9236-2_19.

(6) Kenar, E.; Franken, H.; Forcisi, S.; Wörmann, K.; Häring, H.-U.; Lehmann, R.; Schmitt-Kopplin, P.; Zell, A.; Kohlbacher, O. Automated Label-Free Quantification of Metabolites from Liquid Chromatography–Mass Spectrometry Data. *Mol. Cell. Proteomics MCP* **2014**, *13* (1), 348–359. https://doi.org/10.1074/mcp.M113.031278.

(7) Mahieu, N. G.; Spalding, J. L.; Patti, G. J. Warpgroup: Increased Precision of Metabolomic Data Processing by Consensus Integration Bound Analysis. *Bioinformatics* **2016**, *32* (2), 268–275. https://doi.org/10.1093/bioinformatics/btv564.

(8) Riquelme, G.; Zabalegui, N.; Marchi, P.; Jones, C. M.; Monge, M. E. A Python-Based Pipeline for Preprocessing LC–MS Data for Untargeted Metabolomics Workflows. *Metabolites* **2020**, *10* (10), 416. https://doi.org/10.3390/metabo10100416.

(9) Li, Z.; Lu, Y.; Guo, Y.; Cao, H.; Wang, Q.; Shui, W. Comprehensive Evaluation of Untargeted Metabolomics Data Processing Software in Feature Detection, Quantification and Discriminating Marker Selection. *Anal. Chim. Acta* **2018**, *1029*, 50–57. https://doi.org/10.1016/j.aca.2018.05.001.

(10) Clark, T. N.; Houriet, J.; Vidar, W. S.; Kellogg, J. J.; Todd, D. A.; Cech, N. B.; Linington, R. G. Interlaboratory Comparison of Untargeted Mass Spectrometry Data Uncovers Underlying Causes for Variability. *J. Nat. Prod.* **2021**, *84* (3), 824–835. https://doi.org/10.1021/acs.jnatprod.0c01376.

(11) Yu, H.; Chen, Y.; Huan, T. Computational Variation: An Underinvestigated Quantitative Variability Caused by Automated Data Processing in Untargeted Metabolomics. *Anal. Chem.* **2021**, *93* (25), 8719–8728. https://doi.org/10.1021/acs.analchem.0c03381.

(12) Barupal, D. K.; Baygi, S. F.; Wright, R. O.; Arora, M. Data Processing Thresholds for Abundance and Sparsity and Missed Biological Insights in an Untargeted Chemical Analysis of Blood Specimens for Exposomics. *Front. Public Health* **2021**, *9*, 753. https://doi.org/10.3389/fpubh.2021.653599.

(13) Baran, R. Untargeted Metabolomics Suffers from Incomplete Raw Data Processing. *Metabolomics* **2017**, *13* (9), 107. https://doi.org/10.1007/s11306-017-1246-3.

(14) Hohrenk, L. L.; Itzel, F.; Baetz, N.; Tuerk, J.; Vosough, M.; Schmidt, T. C. Comparison of Software Tools for Liquid Chromatography–High-Resolution Mass Spectrometry Data Processing in Nontarget Screening of Environmental Samples. *Anal. Chem.* **2019**. https://doi.org/10.1021/acs.analchem.9b04095.

(15) Baker, E. S.; Patti, G. J. Perspectives on Data Analysis in Metabolomics: Points of Agreement and Disagreement from the 2018 ASMS Fall Workshop. *J. Am. Soc. Mass Spectrom.* **2019**, *30* (10), 2031–2036. https://doi.org/10.1007/s13361-019-02295-3.

(16) Guo, J.; Shen, S.; Xing, S.; Chen, Y.; Chen, F.; Porter, E. M.; Yu, H.; Huan, T. EVA: Evaluation of Metabolic Feature Fidelity Using a Deep Learning Model Trained With Over 25000 Extracted Ion Chromatograms. *Anal. Chem.* **2021**. https://doi.org/10.1021/acs.analchem.1c01309.

(17) Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; Magnes, C. IPO: A Tool for Automated Optimization of XCMS Parameters. *BMC Bioinformatics* **2015**, *16* (1), 118. https://doi.org/10.1186/s12859-015-0562-8.

(18) Albóniga, O. E.; González, O.; Alonso, R. M.; Xu, Y.; Goodacre, R. Optimization of XCMS Parameters for LC–MS Metabolomics: An Assessment of Automated versus Manual Tuning and Its Effect on the Final Results. *Metabolomics* **2020**, *16* (1), 14. https://doi.org/10.1007/s11306-020-1636-9.

(19) Hao, L.; Wang, J.; Page, D.; Asthana, S.; Zetterberg, H.; Carlsson, C.; Okonkwo, O. C.; Li, L. Comparative Evaluation of MS-Based Metabolomics Software and Its Application to Preclinical Alzheimer's Disease. *Sci. Rep.* **2018**, *8* (1), 9291. https://doi.org/10.1038/s41598-018-27031-x.

(20) Hemmer, S.; Manier, S. K.; Fischmann, S.; Westphal, F.; Wagmann, L.; Meyer, M. R. Comparison of Three Untargeted Data Processing Workflows for Evaluating LC-HRMS Metabolomics Data. *Metabolites* **2020**, *10* (9), 378. https://doi.org/10.3390/metabo10090378.

(21) Chen, Y.; Xu, J.; Zhang, R.; Shen, G.; Song, Y.; Sun, J.; He, J.; Zhan, Q.; Abliz, Z. Assessment of Data Pre-Processing Methods for LC-MS/MS-Based Metabolomics of Uterine Cervix Cancer. *Analyst* **2013**, *138* (9), 2669–2677. https://doi.org/10.1039/C3AN36818A.

(22) Chaker, J.; Gilles, E.; Léger, T.; Jégou, B.; David, A. From Metabolomics to HRMS-Based Exposomics: Adapting Peak Picking and Developing Scoring for MS1 Suspect Screening. *Anal. Chem.*

2021, *93* (3), 1792–1800. https://doi.org/10.1021/acs.analchem.0c04660.

(23) Yan, B.-P.; Cao, C.-M.; Hou, J.-J.; Bi, Q.-R.; Yang, M.; Qi, P.; Shi, X.-J.; Wang, J.-W.; Wu, W.-Y.; Guo, D.-A. With Guide of Spike-in Experiment for Optimizing Workflow of LC-MS Data Processing in Metabolomics. *Nat. Prod. Commun.* **2017**, *12* (8), 1934578X1701200837. https://doi.org/10.1177/1934578X1701200837.

(24) Müller, E.; Huber, C.; Beckers, L.-M.; Brack, W.; Krauss, M.; Schulze, T. A Data Set of 255,000 Randomly Selected and Manually Classified Extracted Ion Chromatograms for Evaluation of Peak Detection Methods. *Metabolites* **2020**, *10* (4), 162. https://doi.org/10.3390/metabo10040162.

(25) El Abiead, Y.; Milford, M.; Salek, R. M.; Koellensperger, G. MzRAPP: A Tool for Reliability Assessment of Data Pre-Processing in Non-Targeted Metabolomics. *Bioinformatics* **2021**, No. btab231. https://doi.org/10.1093/bioinformatics/btab231.

(26) Pang, Z.; Chong, J.; Li, S.; Xia, J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. *Metabolites* **2020**, *10* (5), 186. https://doi.org/10.3390/metabo10050186.

(27) Haug, K.; Cochrane, K.; Nainala, V. C.; Williams, M.; Chang, J.; Jayaseelan, K. V.; O'Donovan, C. MetaboLights: A Resource Evolving in Response to the Needs of Its Scientific Community. *Nucleic Acids Res.* **2020**, *48* (D1), D440–D444. https://doi.org/10.1093/nar/gkz1019.

(28) Schoeny, H.; Rampler, E.; Abiead, Y. E.; Hildebrand, F.; Zach, O.; Hermann, G.; Koellensperger, G. A Combined Flow Injection/Reversed-Phase Chromatography–High-Resolution Mass Spectrometry Workflow for Accurate Absolute Lipid Quantification with 13C Internal Standards. *Analyst* **2021**, *146* (8), 2591–2599. https://doi.org/10.1039/D0AN02443K.

(29) Kovářová, J.; Nagar, R.; Faria, J.; Ferguson, M. A. J.; Barrett, M. P.; Horn, D. Gluconeogenesis Using Glycerol as a Substrate in Bloodstream-Form Trypanosoma Brucei. *PLoS Pathog.* **2018**, *14* (12), e1007475. https://doi.org/10.1371/journal.ppat.1007475.

(30) Chaleckis, R.; Murakami, I.; Takada, J.; Kondoh, H.; Yanagida, M. Individual Variability in Human Blood Metabolites Identifies Age-Related Differences. *Proc. Natl. Acad. Sci.* **2016**, *113* (16), 4252–4259. https://doi.org/10.1073/pnas.1603023113.

(31) Rusz, M.; Del Favero, G.; El Abiead, Y.; Gerner, C.; Keppler, B. K.; Jakupec, M. A.; Koellensperger, G. Morpho-metabotyping the Oxidative Stress Response. *Sci. Rep.* **2021**, *11* (1), 15471. https://doi.org/10.1038/s41598-021-94585-8.

(32) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A Cross-Platform Toolkit for Mass Spectrometry and Proteomics. *Nat. Biotechnol.* **2012**, *30* (10), 918–920. https://doi.org/10.1038/nbt.2377.

(33) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: An Open Source Document Editor for Creating and Analyzing Targeted Proteomics Experiments. *Bioinformatics* **2010**, *26* (7), 966–968. https://doi.org/10.1093/bioinformatics/btq054.

(34) McLean, C.; Kujawinski, E. B. AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing. *Anal. Chem.* **2020**, *92* (8), 5724–5732. https://doi.org/10.1021/acs.analchem.9b04804.

(35) Delabriere, A.; Warmer, P.; Brennsteiner, V.; Zamboni, N. SLAW: A Scalable and Self-Optimizing Processing Workflow for Untargeted LC-MS (Submitted), 2021.

(36) Tautenhahn, R.; Böttcher, C.; Neumann, S. Highly Sensitive Feature Detection for High Resolution LC/MS. *BMC Bioinformatics* **2008**, *9* (1), 504. https://doi.org/10.1186/1471-2105-9-504.

(37) Myers, O. D.; Sumner, S. J.; Li, S.; Barnes, S.; Du, X. 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), 8696–8703. https://doi.org/10.1021/acs.analchem.7b00947.

(38) Gloaguen, Y.; Kirwan, J.; Beule, D. *Deep Learning Assisted Peak Curation for Large Scale LC-MS Metabolomics*; 2020; p 2020.08.09.242727. https://doi.org/10.1101/2020.08.09.242727.

(39) Zhang, W.; Zhao, P. X. Quality Evaluation of Extracted Ion Chromatograms and Chromatographic Peaks in Liquid Chromatography/Mass Spectrometry-Based Metabolomics Data. *BMC Bioinformatics* **2014**, *15* (11), S5. https://doi.org/10.1186/1471-2105-15-S11-S5.

(40) Choi, D. Y.; Row, K. H. Theoretical Analysis of Chromatographic Peak Asymmetry and Sharpness by the Moment Method Using Two Peptides. *Biotechnol. Bioprocess Eng.* 2004, *9* (6), 495–499. https://doi.org/10.1007/BF02933492.