mzrtsim: Raw Data Simulation for Reproducible Gas/Liquid Chromatography–Mass Spectrometry Based Non-targeted Metabolomics Data Analysis

Miao Yu[1*], Vivek Philip[2]

[1] Affiliation 1; The Jackson laboratory, 10 Discovery Drive Farmington, CT USA 06032
[2] Affiliation 2; The Jackson laboratory, 600 Main Street, Bar Harbor, ME USA 04609
* Correspondence: miao.yu@jax.org

Abstract

Reproducibility of data analysis is pivotal in the context of non-targeted metabolomics based on mass spectrometry. While various algorithms have been proposed for feature extraction, their validation often revolves around a limited set of known compounds or standards.While data simulation is widely used in other omics studies, many simulations are focused on the feature level, neglecting uncertainties inherent in the feature extraction process for metabolomics mass spectrometry data. In this study, we introduce a R package called 'mzrtsim' to simulate MS1 full scan raw data in the mzML format. Unlike simulations solely based on statistical distributions, our approach leverages experimental data from the MassBank of North America(MoNA) and the human metabolome database(HMDB). We also developed algorithms to realistically simulate chromatography peaks, accounting for factors such as tailing and leading. The results of our study demonstrate the potential of this tool not only for comparing established metabolomics software (e.g., XCMS, mzMine, and OpenMS) against a ground truth, but also for evaluating their performance under challenging chromatographic conditions often encountered in non-targeted metabolomics. We found that the investigated software all introduced large amounts of false positive peaks (more than 75% of detected peaks by default setting) and loss compounds with less peaks. They also showed different sensitivity to trailing and leading peaks and we also found setting appropriate intensity cutoffs will reduce the influences from redundant peaks. This R package is free and available online(https://github.com/yufree/mzrtsim).

Introduction

The primary objective of metabolomics is the comprehensive identification and precise quantification of metabolites within a biological system[1]. To conduct metabolomics data analysis, researchers are required to fine-tune a multitude of software, settings, and parameters during the bioinformatic processing of raw data[2,3]. It is important to note that these adjustments do not occur in isolation, frequently leading to a cascade of repercussions across various facets of the experiment[4,5].

In mass spectrometry based metabolomics research, numerous software has been proposed to process the raw data and extract features[6–8]. However, past studies have illuminated significant disparities in the performance of these software tools and algorithmic parameters[9]. For instance, when applied to identical datasets, five distinct software packages yield comparable true positive peaks, yet exhibit variations in the relative quantification of true features. Furthermore, the task of evaluating these software solutions or algorithms is consistently challenging, given the multitude of parameters involved and the inherent uncertainties in non-targeted analysis. Previous approaches have often relied on the utilization of known compound standards[5], publicly available datasets[9] or manually checking[10] of peak shape to choose software, which might not be reproducible for other researchers. In this context, a “white-box” and reproducible compounds will be helpful to assess software performance or facilitate developmental efforts for metabolomics data analysis.

Proteomics and lipidomics research have used simulation data with a known ground truth[11,12].Typically, proteomics and lipidomics studies involve data-dependent acquisition (DDA) or data-independent acquisition (DIA) methods for qualitative assessment. These modes have also found application in metabolomics, particularly with pooled samples. However, it's important to note that these modes can result in the loss of MS1 data coverage[13]. The regular workflow in metabolomics begins with full-scan mode for the quantitative analysis, aiming to identify potential biomarkers. Subsequently, MS/MS spectra of selected ions are extracted for annotation purposes. Consequently, the majority of metabolomics software tools are primarily designed to extract MS1 peaks or features[14–17]. In this case, simulation of MS1 data for metabolomics will be helpful to evaluate metabolomics software.

Prior efforts in simulating metabolomics data have predominantly concentrated on the compound-level data representation[18,19]. However, real experimental data frequently encompass redundant peaks, offering a glimpse into the intricate nature of

metabolomics datasets[20]. Many of these redundant peaks, including isotopologues, adducts, fragmental ions and multiple charged ions, often exhibit correlated intensities with base peaks and are susceptible to the influence of background noise with lower intensity[21]. Such redundant peaks might be useful for downstream annotation analysis[22]. Therefore, in the context of simulation, it becomes imperative to account for the presence of noisy background signals and incorporate the generation of redundant peaks that parallel the characteristics observed in actual experimental data.

On the other hand, the performance of chromatography plays a pivotal role, and previous studies have often centered on assessing software based on chromatography peak quality[23,24]. Nevertheless, achieving consistent peak quality in real samples can be a challenging endeavor considering sample preparation and instrument analysis. Unlike targeted analysis where known ions are optimized, untargeted analysis doesn't have specific compounds to be optimized. Consequently, issues with peak quality can arise, particularly when dealing with unforeseen compounds. These fluctuations in peak quality might impact the effectiveness of peak picking algorithms. In this context, the inclusion of direct simulation of peak shapes, encompassing characteristics like leading and tailing peaks, can prove highly beneficial. Such simulations would not only provide a more comprehensive evaluation of software performance but also help account for the variability in peak quality encountered in real-world untargeted metabolomics studies.

In this research endeavor, we present the 'mzrtsim' package, designed specifically for the generation of mzML files with the primary aim of evaluating metabolomics software. This tool is equipped to produce simulated experimental data that faithfully replicates the issues encountered in real metabolomics datasets. These include the incorporation of redundant peaks and background noise, mirroring the complexity of actual experimental data. Additionally, 'mzrtsim' is uniquely capable of simulating chromatography peak quality issues with reference to known ground truth, providing a comprehensive framework for assessing and enhancing the performance of metabolomics software.

Methods

The experimental full scan spectrum data from MassBank of North America(MoNA)[25] for liquid chromatography-mass spectrometry (LC-MS) and the human metabolome database(HMDB)[26] for gas chromatography-mass spectrometry (GC-MS) was extracted for simulation purpose with redundant peaks. In brief, MoNA database contains 6075 experimental compounds spectra from 1926 unique compounds. 2968 compound spectra from 1114 compounds are collected on high resolution mass

spectrometry. Among those spectra, the average peak number is 43.5 and the median peak number is 15. As for HMDB, 10096 experimental spectra are collected from 2295 unique compounds. The average peak number for HMDB is 87.6 and the median peak number is 48. Those experimental data contain not only molecular ions, but also redundant peaks and background noises.

For each compound's spectra, a signal to noise('sn') ratio could be set for the base peaks. For example, a compound has two ions with one base peak and another isotopologue peak of the base peak. The intensity of the isotopologue peak is 5% of the base peaks. Then, the 'sn' of 100 with 'baseline' 100 for this compound means the base peak of the compound should have 10000 as intensity and 500 for the isotopologue peak.

Each ion in one spectrum will generate its own chromatography peaks by user defined scan rate and peak width. For example, a scan rate of 0.2 spectra per seconds with 10s peak width will simulate 50 spectra of one compound. Without tailing, each ion in the mass spectra should follow a normal distribution, considering the relative intensity within each spectra. Normal distribution parameters could be set up with mean as 'baseline' and standard deviation as 'baselinesd' for each ion's chromatography peaks. The retention time could also be set for certain compounds. Meanwhile, peak quality could be set by 'tailingfactor' for the simulation and the users could simulate tailing peaks('tailingfactor' >1) and leading peaks('tailingfactor' <1) for certain compounds.

With simulation of single spectrum and chromatography peaks, mzrtsim package[27] could generate an mzML file through the Spectra package[28]. With control of peak width for base peaks and retention times, users could generate data sets for within groups and across group samples. When the peak width is a single value, the simulated peak width will be from a poisson distribution with this value as lambda. The output files should have mzML files and csv files with mass to charge ratio, retention time and compound name for each peak. Such csv files will be used to evaluate the performance of different software.

With mzrtsim package (see figure 1), we could benchmark different metabolomics software. We simulate the same data sets to compare XCMS[6], OpenMS[8], and MZmine 3[29]. We also test their permanence on different types of chromatography peaks. Then we tested the influence of redundant peaks with lower intensity on XCMS. The overlapped peaks are defined as peaks with mass to charge ratio shift within 5 ppm and retention time shift within 5 seconds. All the simulated data are available online.

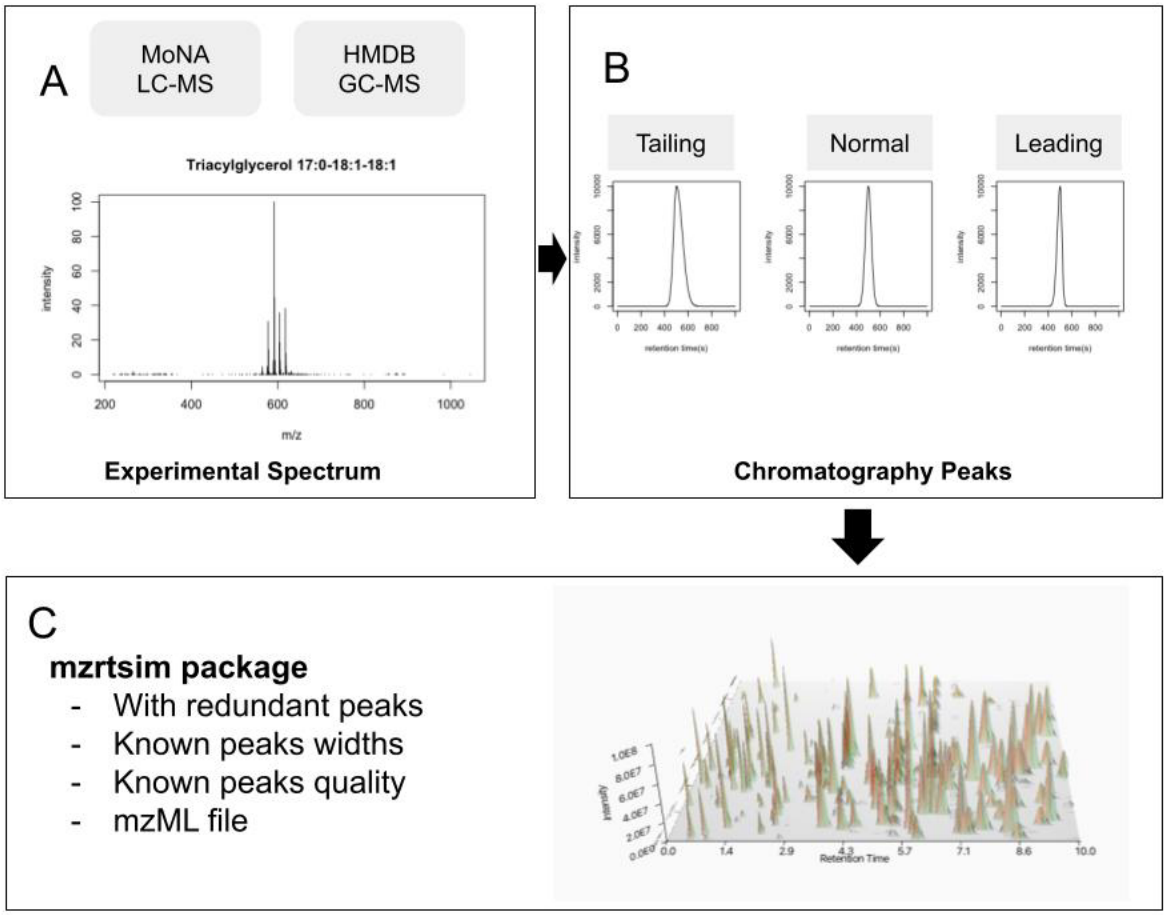

Figure1. mzML simulation process of mzrtsim package. (A)Experimental spectrum from publicly available databases(MoNA and HMDB) will be used for simulation and (B) each ion in the spectrum will generate chromatography peaks to (C) output the final mzML file with known information for each peak.

Results and Discussion

Software comparison for peak picking

We simulated 10 mzML files with 100 compounds' high resolution spectra characterized by peaks following a normal distribution. Those 100 compounds generated 593 detectable chromatography peaks. Then peak picking was performed using XCMS, OpenMS, and MZmine 3 with default settings. Results showed that those three software found 41 compounds with slightly different profiles. However, their peak profiles are quite different. XCMS found 495 peaks and 121 peaks are real peaks. OpenMS found 926 peaks and 179 peaks are from simulated compounds. MZmine 3 detected 776 peaks with 186 true positive peaks. More than 75% of the detected peaks

are false positive peaks in all of those three software. However, a closer examination using a Venn diagram (figure 2) revealed that only 105 peaks were shared among the true positives. This finding suggests that while the software tools exhibited varying performance in detecting false positive peaks, there was a substantial overlap in their identification of true peaks. Therefore, in practical applications, employing multiple software tools for peak extraction and subsequently focusing on the overlapping peaks is advisable to mitigate the impact of false positives and optimize analysis efficiency.

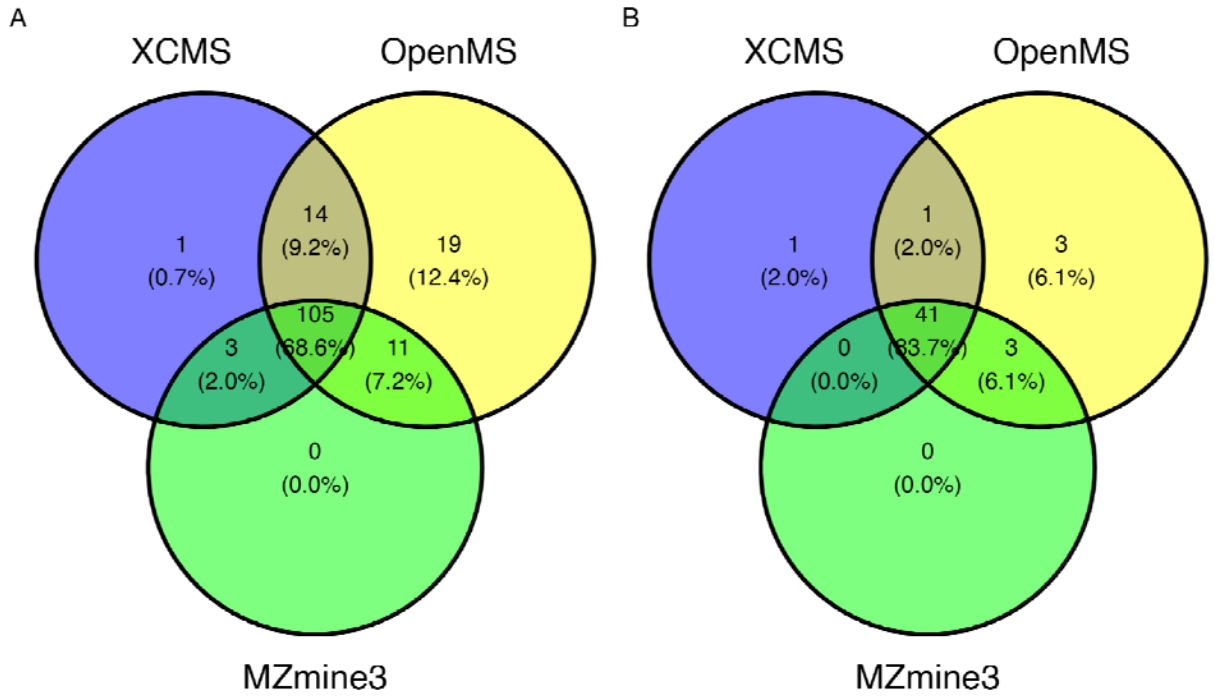

Figure 2. Venn diagram for real detected peaks(left) and detected compounds(right) among XCMS, OpenMS, and MZmine 3.

Since more than half of the compounds are missing for all those three compounds, we also checked the missed compounds. Only 75 missing peaks are not captured by peak picking software while they are from 51 unique compounds. In this case, when the compounds generate less peaks, the software might miss them.

Software comparison for tailing/leading peaks

We also simulate the tailing and leading chromatography peaks by changing the tailing factor of all peaks. As shown in table 1, different software also displays different

sensitivity. XCMS's performance is quite stable across different types of peaks while finding the minimal numbers of true peaks. MZmine also showed a stable performance with a similar false positive rate(75%-77%) compared with XCMS. OpenMS, on the other hand, finds more false positive peaks in leading peaks than tailing peaks. In this case, for data with known leading or tailing peaks, it's better to use MZmine 3 for peak picking.

Table 1. Extracted peaks/True positive peaks/Detected compounds of XCMS, OpenMS, MZmine 3 peak picking algorithms for 100 simulated compounds with 593 normal, leading and tailing chromatography peaks.

	XCMS	OpenMS	MZmine 3
normal	495/121/43	926/179/48	776/186/44
leading	495/122/43	1182/188/51	789/185/44
tailing	489/120/43	772/159/48	787/193/45

Redundant peaks' influences on statistical analysis

Redundant peaks, characterized by their lack of new biological information and interdependence with other peaks from the same compounds, can affect the efficiency of data analysis. To address this, we conducted experiments involving two groups of samples, each containing known compounds with altered intensities. Importantly, we applied different intensity cutoffs to each set of spectra, with the rationale of reducing the volume of real data for expedited high-throughput analysis.

In a scenario where 30 out of 100 peaks exhibited changes between the two groups of files, we employed intensity cutoffs of 0 and 0.05 to examine their impact on statistical analysis. When the cutoff is set at 0, all peaks derived from the same compounds were retained in the analysis. However, when the cutoff was increased to 0.05, peaks with relative intensities lower than 5% of the base peaks were filtered out. This adjustment resulted in a notable reduction in the typical file size of mzML files, shrinking from 185MB to 34.5MB for the same compounds.

Specifically, using a cutoff of 0, we identified 3,400 real peaks from 100 compounds, whereas with a cutoff of 0.05, only 533 peaks remained. Subsequently, employing XCMS for peak extraction, we obtained 2,546 peaks for cutoff 0 and 496 peaks for cutoff 0.05. Among these, 914 peaks from 26 compounds exhibited changes in files

with a cutoff of 0, and 817 peaks were significant according to t-tests with Benjamini-Hochberg procedure[30] to control false discovery rate, with 15 compounds classified as false positives. For the cutoff of 0.05, 154 peaks changed from 27 compounds, and 154 peaks were significant using t-tests with false discovery control, with 8 compounds identified as false positives.

This analysis demonstrates that an increase in redundant peaks leads to a higher occurrence of false positives for differential analysis. In practical applications, a suitable cutoff to remove redundant peaks is crucial to avoid bias on compounds with more peaks.

Other potential applications

In addition to comparing various peak-picking algorithms, the simulation of mzML files holds significant potential for enhancing existing algorithms in a broader context within the field of metabolomics. For instance, researchers can utilize this methodology to introduce matrix ions as a form of background noise into their data. This allows for the investigation of how these additional ions may impact subsequent statistical analyses, providing valuable insights into data robustness and the potential need for noise mitigation strategies.Furthermore, when dealing with known compounds with predefined retention times, mzML file simulation offers a unique opportunity to assess the influence of co-elution effects and evaluate the effectiveness of peak alignment algorithms. These insights are crucial for refining data processing pipelines and improving the accuracy of biomarker discovery. Beyond these applications, mzML file simulation can also serve as a valuable tool for validating emerging data analysis methods, such as those based on deep learning techniques and image analysis. By generating controlled datasets with known properties, researchers can rigorously test and fine-tune these innovative approaches, ultimately advancing the state-of-the-art in metabolomics data analysis.

Conclusion

In summary, we have introduced a dedicated raw file simulation tool designed specifically for chromatography coupled with mass spectrometry data. Our initial findings have unveiled an intriguing insight into the performance of widely-used software in this domain – the presence of false positive peaks. This revelation underscores the importance of amalgamating their results to achieve more robust outcomes. Additionally, our investigation has shed light on the role of redundant peaks,

which can introduce false positives into the analysis. the mzrtsim package highlights critical insights into the challenges and opportunities for metabolomics data analysis software development.

Author Contributions

M. Y.: Conceptualization, software development, data curation, visualization, writing-original draft, writing-review and editing; V.P.: writing review and editing, supervision. All authors read, reviewed, and accepted the final manuscript.

Data Availability Statement

All the data associated with the present study can be accessed on Zenodo (https://doi.org/10.5281/zenodo.8299782) and software could be downloaded from GitHub (https://github.com/yufree/mzrtsim)

Conflicts of Interest

The authors declare no conflict of interests. 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.

References

1. González-Riano, C.; Dudzik, D.; Garcia, A.; Gil-de-la-Fuente, A.; Gradillas, A.; Godzien, J.; López-Gonzálvez, Á.; Rey-Stolle, F.; Rojo, D.; Ruperez, F.J.; et al. Recent Developments along the Analytical Process for Metabolomics Workflows. *Anal. Chem.* **2020**, *92*, 203–226, doi:10.1021/acs.analchem.9b04553.
2. Myers, O.D.; Sumner, S.J.; Li, S.; Barnes, S.; Du, X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. *Anal. Chem.* **2017**, *89*, 8689–8695, doi:10.1021/acs.analchem.7b01069.
3. Weber, R.J.M.; Lawson, T.N.; Salek, R.M.; Ebbels, T.M.D.; Glen, R.C.; Goodacre, R.; Griffin, J.L.; Haug, K.; Koulman, A.; Moreno, P.; et al. Computational Tools and Workflows in Metabolomics: An International Survey Highlights the Opportunity for Harmonisation through Galaxy. *Metabolomics* **2017**, *13*, doi:10.1007/s11306-016-1147-x.
4. Liao, J.; Zhang, Y.; Zhang, W.; Zeng, Y.; Zhao, J.; Zhang, J.; Yao, T.; Li, H.; Shen, X.; Wu, G.; et al. Different Software Processing Affects the Peak Picking and Metabolic Pathway Recognition of Metabolomics Data. *J. Chromatogr. A* **2023**, *1687*, 463700, doi:10.1016/j.chroma.2022.463700.

5. 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, doi:10.1016/j.aca.2018.05.001.
6. Domingo-Almenara, X.; Siuzdak, G. Metabolomics Data Processing Using XCMS. In *Computational Methods and Data Analysis for Metabolomics*; Li, S., Ed.; Methods in Molecular Biology; Springer US: New York, NY, 2020; pp. 11–24 ISBN 978-1-07-160239-3.
7. Korf, A.; Fouquet, T.; Schmid, R.; Hayen, H.; Hagenhoff, S. Expanding the Kendrick Mass Plot Toolbox in MZmine 2 to Enable Rapid Polymer Characterization in Liquid Chromatography−Mass Spectrometry Data Sets. *Anal. Chem.* **2020**, *92*, 628–633, doi:10.1021/acs.analchem.9b03863.
8. Röst, H.L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weisser, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.-C.; Gutenbrunner, P.; Kenar, E.; et al. OpenMS: A Flexible Open-Source Software Platform for Mass Spectrometry Data Analysis. *Nat. Methods* **2016**, *13*, 741–748, doi:10.1038/nmeth.3959.
9. El Abiead, Y.; Milford, M.; Schoeny, H.; Rusz, M.; Salek, R.M.; Koellensperger, G. Power of MzRAPP-Based Performance Assessments in MS1-Based Nontargeted Feature Detection. *Anal. Chem.* **2022**, *94*, 8588–8595, doi:10.1021/acs.analchem.1c05270.
10. Chetnik, K.; Petrick, L.; Pandey, G. MetaClean: A Machine Learning-Based Classifier for Reduced False Positive Peak Detection in Untargeted LC–MS Metabolomics Data. *Metabolomics Off. J. Metabolomic Soc.* **2020**, *16*, 117, doi:10.1007/s11306-020-01738-3.
11. Hutchins, P.D.; Russell, J.D.; Coon, J.J. Accelerating Lipidomic Method Development through in Silico Simulation. *Anal. Chem.* **2019**, *91*, 9698–9706, doi:10.1021/acs.analchem.9b01234.
12. Leeming, M.G.; Ang, C.-S.; Nie, S.; Varshney, S.; Williamson, N.A. Simulation of Mass Spectrometry-Based Proteomics Data with Synthedia. *Bioinforma. Adv.* **2023**, *3*, vbac096, doi:10.1093/bioadv/vbac096.
13. Yu, M.; Dolios, G.; Petrick, L. Reproducible Untargeted Metabolomics Workflow for Exhaustive MS2 Data Acquisition of MS1 Features. *J. Cheminformatics* **2022**, *14*, 6, doi:10.1186/s13321-022-00586-8.
14. Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; et al. IPO: A Tool for Automated Optimization of XCMS Parameters. *BMC Bioinformatics* **2015**, *16*, 118, doi:10.1186/s12859-015-0562-8.
15. Delabriere, A.; Warmer, P.; Brennsteiner, V.; Zamboni, N. SLAW: A Scalable and Self-Optimizing Processing Workflow for Untargeted LC-MS. *Anal. Chem.* **2021**, *93*, 15024–15032, doi:10.1021/acs.analchem.1c02687.
16. McLean, C.; Kujawinski, E.B. AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing. *Anal. Chem.* **2020**, *92*, 5724–5732, doi:10.1021/acs.analchem.9b04804.
17. Ni, Y.; Su, M.; Qiu, Y.; Jia, W.; Du, X. ADAP-GC 3.0: Improved Peak Detection and Deconvolution of Co-Eluting Metabolites from GC/TOF-MS Data for Metabolomics Studies. *Anal. Chem.* **2016**, *88*, 8802–8811, doi:10.1021/acs.analchem.6b02222.
18. Manimaran, S.; Selby, H.M.; Okrah, K.; Ruberman, C.; Leek, J.T.; Quackenbush, J.; Haibe-Kains, B.; Bravo, H.C.; Johnson, W.E. BatchQC: Interactive Software for Evaluating Sample and Batch Effects in Genomic Data. *Bioinformatics* **2016**, *32*, 3836–3838, doi:10.1093/bioinformatics/btw538.
19. Mendes, P.; Camacho, D.; de la Fuente, A. Modelling and Simulation for Metabolomics Data Analysis. *Biochem. Soc. Trans.* **2005**, *33*, 1427–1429, doi:10.1042/BST0331427.
20. Mahieu, N.G.; Patti, G.J. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Anal. Chem.* **2017**, *89*, 10397–10406, doi:10.1021/acs.analchem.7b02380.

21. Yu, M.; Olkowicz, M.; Pawliszyn, J. Structure/Reaction Directed Analysis for LC-MS Based Untargeted Analysis. *Anal. Chim. Acta* **2019**, *1050*, 16–24, doi:10.1016/j.aca.2018.10.062.
22. Chen, L.; Lu, W.; Wang, L.; Xing, X.; Chen, Z.; Teng, X.; Zeng, X.; Muscarella, A.D.; Shen, Y.; Cowan, A.; et al. Metabolite Discovery through Global Annotation of Untargeted Metabolomics Data. *Nat. Methods* **2021**, *18*, 1377–1385, doi:10.1038/s41592-021-01303-3.
23. Gloaguen, Y.; Kirwan, J.A.; Beule, D. Deep Learning-Assisted Peak Curation for Large-Scale LC-MS Metabolomics. *Anal. Chem.* **2022**, *94*, 4930–4937, doi:10.1021/acs.analchem.1c02220.
24. Guo, J.; Shen, S.; Huan, T. Paramounter: Direct Measurement of Universal Parameters To Process Metabolomics Data in a "White Box." *Anal. Chem.* **2022**, doi:10.1021/acs.analchem.1c04758.
25. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; et al. MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. *J. Mass Spectrom.* **2010**, *45*, 703–714, doi:https://doi.org/10.1002/jms.1777.
26. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617, doi:10.1093/nar/gkx1089.
27. Yu, M.; Roszkowska, A.; Pawliszyn, J. Simulation-Based Comprehensive Study of Batch Effects in Metabolomics Studies 2019, 2019.12.16.878637.
28. Rainer, J.; Vicini, A.; Salzer, L.; Stanstrup, J.; Badia, J.M.; Neumann, S.; Stravs, M.A.; Verri Hernandes, V.; Gatto, L.; Gibb, S.; et al. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R. *Metabolites* **2022**, *12*, 173, doi:10.3390/metabo12020173.
29. Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T.S.; Bushuiev, R.; Murray, K.J.; Hoffmann, N.; Lu, M.; et al. Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3. *Nat. Biotechnol.* **2023**, *41*, 447–449, doi:10.1038/s41587-023-01690-2.
30. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300.