analytical chemistry

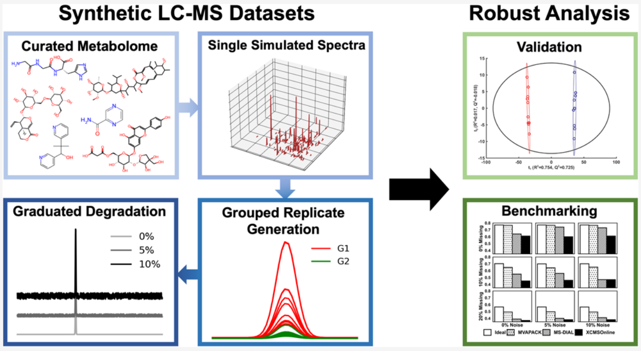
pubs.acs.org/ac Article

Simulated LC−MS Data Set for Assessing the Metabolomics Data Processing Pipeline Implemented into MVAPACK

Christopher P. Jurich, Micah J. Jeppesen, Isin T. Sakallioglu, Aline De Lima Leite, Joseph D. Yesselman,* and Robert Powers*

Cite This: *Anal. Chem.* 2024, 96, 12943−12956 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



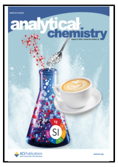
ABSTRACT: Metabolomics commonly relies on using one-dimensional (1D) ^{1}H NMR spectroscopy or liquid chromatography−mass spectrometry (LC−MS) to derive scientific insights from large collections of biological samples. NMR and MS approaches to metabolomics require, among other issues, a data processing pipeline. Quantitative assessment of the performance of these software platforms is challenged by a lack of standardized data sets with "known" outcomes. To resolve this issue, we created a novel simulated LC−MS data set with known peak locations and intensities, defined metabolite differences between groups (i.e., fold change > 2, coefficient of variation ≤ 25%), and different amounts of added Gaussian noise (0, 5, or 10%) and missing features (0, 10, or 20%). This data set was developed to improve benchmarking of existing LC−MS metabolomics software and to validate the updated version of our MVAPACK software, which added gas chromatography−MS and LC−MS functionality to its existing 1D and two-dimensional NMR data processing capabilities. We also included two experimental LC−MS data sets acquired from a standard mixture and*Mycobacterium smegmatis*cell lysates since a simulated data set alone may not capture all the unique characteristics and variability of real spectra needed to assess software performance properly. Our simulated and experimental LC−MS data sets were processed with the MS-DIAL and XCMSOnline software packages and our MVAPACK toolkit to showcase the utility of our data sets to benchmark MVAPACK against community standards. Our results demonstrate the enhanced objectivity and clarity of software assessment that can be achieved when both simulated and experimental data are employed since distinctly different software performances were observed with the simulated and experimental LC−MS data sets. We also demonstrate that the performance of MVAPACK is equivalent to or exceeds existing LC−MS software programs while providing a single platform for processing and analyzing both NMR and MS data sets.

INTRODUCTION

Metabolomics is the quantitative analysis of small molecule metabolites present in biological samples.[1] These metabolites include cofactors, substrates, and end products of enzymatic reactions, which play pivotal roles in cellular processes and are found in diverse tissues, cell lysates, and biofluids such as serum, saliva, and urine.[2−4] Collectively, these metabolites constitute an organism's metabolome, which can comprise thousands of distinct compounds.[5] Given the sheer diversity and abundance of metabolites in various biological samples, the metabolome serves as a robust proxy for understanding a wide variety of biological processes and disease states.[6] For instance, the abundance of specific metabolite biomarkers has been correlated with Parkinson's disease, oral cancers, diabetes, heart disease, and an assortment of other human ailments.[7−10] Thus, quantifying metabolic biomarkers is critical for deciphering disease mechanisms, formulating diagnostic and prognostic tests, and propelling drug discovery and development.[11−15] Achieving these and other laudable goals is critically dependent on the complete and thorough characterization of the metabolome for a biological system. However, realizing these objectives hinges on comprehensively characterizing an organism's metabolome. A significant challenge in this endeavor is identifying and quantifying every detectable metabolite in a sample. This task demands a synergy of advanced analytical techniques and specialized software for data processing and statistical analysis.

Liquid chromatography−mass spectrometry (LC−MS) is the most widely used analytical platform for metabolomics studies because of its high sensitivity and rapid throughput.[16−18] Despite the popularity of LC−MS, spectral

Received: November 3, 2023
Revised: July 5, 2024
Accepted: July 10, 2024
Published: July 30, 2024



https://doi.org/10.1021/acs.analchem.3c04979

Table 1. Summary of Raw Simulated Datasets[a]

data set name	feature count[b]	significant features[b]	non-significant features[b]	median FC	mean % CV (%)	% noise (%)	number missing features[c]	% missing features[c] (%)
SD1	2673	935	1738	8.32	31.42	0	0	0
SD2	2673	935	1738	8.32	31.42	5	0	0
SD3	2673	935	1738	8.32	31.42	10	0	0
SD4	2673	935	1738	8.43	40.09	0	1870	10
SD5	2673	935	1738	8.26	40.06	5	1870	10
SD6	2673	935	1738	8.40	40.09	10	1870	10
SD7	2673	935	1738	8.37	46.81	0	3740	20
SD8	2673	935	1738	8.32	46.81	5	3740	20
SD9	2673	935	1738	8.27	46.81	10	3740	20

[a] Each set contains the same metabolite features, with baseline noise and metabolite features having been added and removed, respectively. The quality of each subsequent data set is degraded such that the quality of SD1 is the highest and the quality of SD9 is the lowest. All statistics are presented for the data sets prior to imputation. Significant compounds are those for which FC ≥ 2.0 and % CV ≤ 25%. [b] Feature counts per replicate in the data set, which comprises 10 replicates for each of the 2 groups (20 replicates in total). [c] Features and percent missing features across the entire data set. Only significant metabolite features are removed from the data set for a total of 18,700 features (935 features × 20 total replicates).

processing, and metabolite annotation follow a complex procedure of spectral alignment, batch correction, peak picking, and metabolite feature grouping, selection, and analysis. Spectral feature selection requires further investigator decision points such as identifying an acceptable minimal fold change (FC ≥ 2−3),[19−21] coefficient of variance (CV < 30%),[19,22] p-values from pairwise comparisons (p < 0.05−0.01),[19,23] importance of each variable (VIP ≥ 1),[24−27] and a missing value threshold (≥80%).[28,29] Further processing of the data matrix requires missing value imputation, normalization, and scaling. The scientific literature contains multiple algorithms for each processing step, creating numerous decision forks and a complex array of possible data processing pipelines dependent on the data type. Consequently, most metabolomics software has only been developed for a single data type, like MS-DIAL[30] and XCMSOnline[31] for LC−MS, which hinders the multiplatform approach needed to improve the coverage of the metabolome.[32]

The diversity and proliferation of metabolomics-related software is a result of the complexity of the data and the data processing steps. Still, it may also result from insufficient guidance from community-certified protocols. For example, despite ongoing efforts,[33−35] benchmark data and community accepted performance standards are lacking, which impedes the further development, assessment, validation, and adoption of a specific data processing pipeline and its underlying software. Benchmarking performance is critical for developers to fine-tune an algorithm's behavior, optimize its outcomes, and for end-users to decide if a given software package fits their analytical needs and expectations and performs reliably. Despite the straightforward utility and need, the metabolomics community currently lacks standardized benchmarks that are universally employed. This is especially problematic for LC−MS software development given the large assortment of instrument vendors, experimental protocols, and chromatography column parameters, in addition to known issues with batch variability, high baseline noise, and missing peaks. These data imperfections originate from biological and technical sources and are independent of resolution and sensitivity.

Existing metabolomics data sets are often utilized with previously established results to benchmark novel software performance. Reanalyzing these data sets with new software can directly compare them with earlier findings. However, this approach is inherently flawed since it assumes the original analysis was correct and complete. Since the data's ground truth is unknown, it is impossible to determine the actual accuracy and precision of a new metabolomics software package. Furthermore, it is unlikely the benchmark data set is broadly accessible by the entire scientific community. More often, separate data sets are used by different groups to evaluate software, which prevents an accurate comparison of performance across groups or software packages. Broadly accessible simulated metabolomics data would directly address these issues and facilitate software development. Simulated data has ground truth, with each peak's waveform, spectral noise, metabolite identity, and concentration fully defined.[5]

The complete knowledge of simulated spectra properties enables a straightforward assessment of algorithmic performance. Spectral features selected by an algorithm can be directly compared to the known composition of metabolites, allowing calculations of false positives (FP), false negatives (FN), sensitivity, selectivity, and other measures of accuracy and precision. Ground truth understanding becomes especially useful when simulated data sets incorporate noise and missing values to mimic the qualities of real spectra. Reducing spectral quality enables stress testing of peak picking, imputation, normalization, and other methods commonly employed in metabolomics data processing. In this manner, simulated data provide feedback on an algorithm's sensitivity to these quality factors.[8] These laudable goals cannot be achieved with experimental data since missing values and noise are not easily tunable factors. While there are clear advantages to simulated data for assessing software performance, there is still inherent value in using experimental data. It is challenging for simulated data sets to capture all the unique characteristics and variability of a real spectrum, especially distributed over a data set comprised of multiple replicate spectra. Instead, as illustrated herein, a preferred approach incorporates both simulated and experimental data in the software evaluation, capitalizing on the strength of both techniques. Despite the clear advantages of using simulated and experimental LC−MS data to assess and validate metabolomics software, the field lacks a data set developed for benchmarking.

In this work, we created LC−MS data sets that combine both simulated and experimental data for algorithm validation and to assess the performance of our MVAPACK toolkit relative to the existing MS-based metabolomics software packages MS-DIAL and XCMSOnline. We have created nine simulated LC−MS data sets with varying quality and noise levels using metabolite feature locations derived from the VIMMS software package https://github.com/glasgowcompbio/vimms[20] and its associated database. We also acquired an experimental LC−MS data set using a standard mixture that captures real spectral quality while minimizing data complexity and simplifying analysis. Both data sets have a set of known peaks that enables an assessment of the performance of the MVAPACK, MS-DIAL, and XCMSOnline software. The LC−MS data set also contains a second experimental LC−MS data set collected on*Mycobacterium smegmatis*cell lysates. This data set captures the realistic characteristics of a typical LC−MS metabolomics experiment and enables an evaluation of relative software performance without ground truth. We anticipate that our simulated and experimental LC−MS data sets will assist and facilitate further software and method development in metabolomics. We also present an updated MVAPACK version to analyze nuclear magnetic resonance (NMR) and LC−MS data. We identified several standard algorithms for LC−MS peak processing from the scientific literature and developed functions to execute these routines within the MVAPACK software platform. In this manner, MVAPACK is the first software toolkit for processing one-dimensional (1D) ^{1}H NMR, two-dimensional (2D) NMR, and LC−MS data sets.

MATERIALS AND METHODS

Summary of Study Design. A synthetic LC−MS data set was created to simulate a metabolomics study of two groups (G1 and G2) with ten replicates per group (Table 1) and 7540 individual peaks with well-defined and high-resolution waveforms. Gaussian curves were fitted to each peak to idealize the data further and enable later manipulation (Figure 1A). Metabolites that differentiate G1 from G2 had a fold change (FC) greater than two and a CV less than or equal to 25% (Figure 1B). Different amounts of added noise (0, 5, or 10%) (Figure 1C) and missing features (0, 10, or 20%) (Figure 1D) were applied to the individual spectra to create a final data set comprising a total of 9 sets of LC−MS spectra. Two additional experimental LC−MS data sets were collected on a Waters (Milford, MA) nanoACQUITY UPLC and XEVO G2-XS QToF system from a standard mixture data set (Table 2) and a biological data set derived from*M. smegmatis*cell lysates. The simulated and experimental LC−MS data sets were then used to validate and benchmark the gas chromatography (GC)/LC−MS data processing pipeline (Table S1) implemented into our MVAPACK[36] metabolomics toolkit (http://bionmr.unl.edu/mvapack.php) and to compare the performance of MVAPACK to MS-DIAL 4.70[30] and XCMSOnline,[31] which are popular metabolomics software packages. Peaks were further filtered by intensity (>100,000 counts) for the standard data set. A detailed description of the experimental protocols can be found in the Supporting Information and on the project's git page (https://git.unl.edu/powers-group/mvapack-lcms-supplemental).

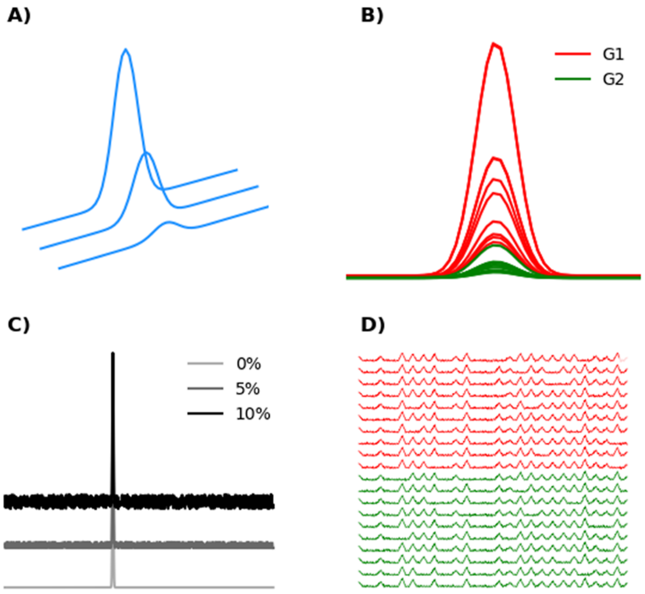


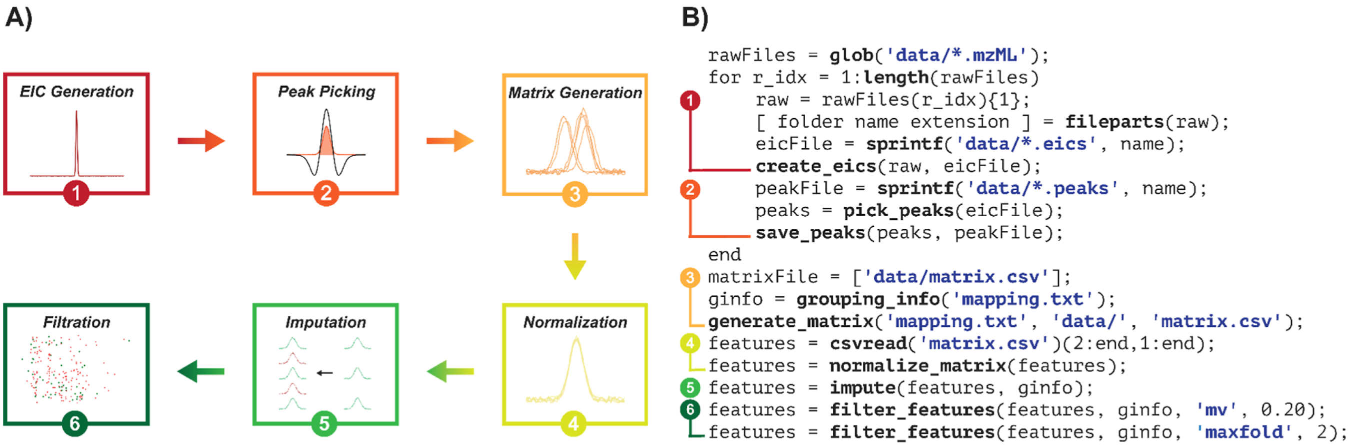
Figure 1. An overview of the construction of the simulated data set. (A) An example of idealized isotope waveforms. (B) The application of statistically significant multipliers to create two groups of metabolite features, G1 (red) and G2 (green), which vary significantly between groups but have low variance within each group. (C) The application of 5% (dark gray) and 10% (black) baseline noise to the original waveform (light gray). (D) Illustration of the removal of features across G1 (red) and G2 (green) replicates to approximate data imperfections common to experimental data sets.

Table 2. Summary of Standard Mixture Dataset

compound	stock (μg/mL)	group 1 (ng/μL)	group 2 (ng/μL)	group 3 (ng/μL)
acetaminophen	10	1	0.5	0.25
caffeine	1.5	0.15	0.075	0.0375
sulfaguanidine	5	0.5	0.25	0.125
sulfadimethoxine	1	0.1	0.05	0.025
val-tyr-val	2.5	0.25	0.125	0.0625
verapamil	0.2	0.02	0.01	0.005
terfenadine	0.2	0.02	0.01	0.005
leucine-enkephalin	2.5	0.25	0.125	0.0625
reserpine	0.6	0.06	0.03	0.015

RESULTS

LC−MS Data Processing Pipeline Added to MVAPACK. LC−MS processing functionality was successfully implemented into MVAPACK, which enables end-to-end analysis of MS-metabolomics data. In keeping with the existing MVAPACK package, LC−MS functionality is modular and provides users with multiple options for each processing step. 40 Octave functions (Table S1) were added to MVAPACK to process LC−MS data sets. The functions comprise the following general capabilities: (i) data input/output (11 functions), (ii) peak alignment (1 function), (iii) feature identification (11 functions), (iv) normalization (4 functions), (v) imputation (4 functions), and (vi) general data processing (9 functions). Acceptable input formats include mzML, mzXML, and proteoWizard.txt formats. Peaks are identified using Gaussian wavelet or Savitzky−Golay filters, and alignment is achieved via an RMSD approach or the ObiWarp method.[31,37] Deisotoping is performed to aggregate the individual peaks into metabolite features. Aggregated matrices are normalized through maximum intensity, quantile, or p-norm methods, and chemically interesting metabolite features are identified through ANOVA, maximum FC, and maximum variance filtration methods. Individual LC−MS processing steps combine to form an analysis pipeline that interfaces with existing MVAPACK modeling functionality. MVAPACK



```
rawFiles = glob('data/*.mzML');
for r_idx = 1:length(rawFiles)
    raw = rawFiles(r_idx){1};
    [ folder name extension ] = fileparts(raw);
    eicFile = sprintf('data/*.eics', name);
    create_eics(raw, eicFile);
    peakFile = sprintf('data/*.peaks', name);
    peaks = pick_peaks(eicFile);
    save_peaks(peaks, peakFile);
end
matrixFile = ['data/matrix.csv'];
ginfo = grouping_info('mapping.txt');
generate_matrix('mapping.txt', 'data/', 'matrix.csv');
features = csvread('matrix.csv')(2:end,1:end);
features = normalize_matrix(features);
features = impute(features, ginfo);
features = filter_features(features, ginfo, 'mv', 0.20);
features = filter_features(features, ginfo, 'maxfold', 2);
```

Figure 2. An overview of the MVAPACK LC−MS processing workflow (A) generic workflow for LC−MS spectra data processing and the corresponding (B) MVAPACK Octave commands. Each processing step is handled with only a few lines of octave code and the example script can handle a large volume of experimental replicates (see Supporting Information).

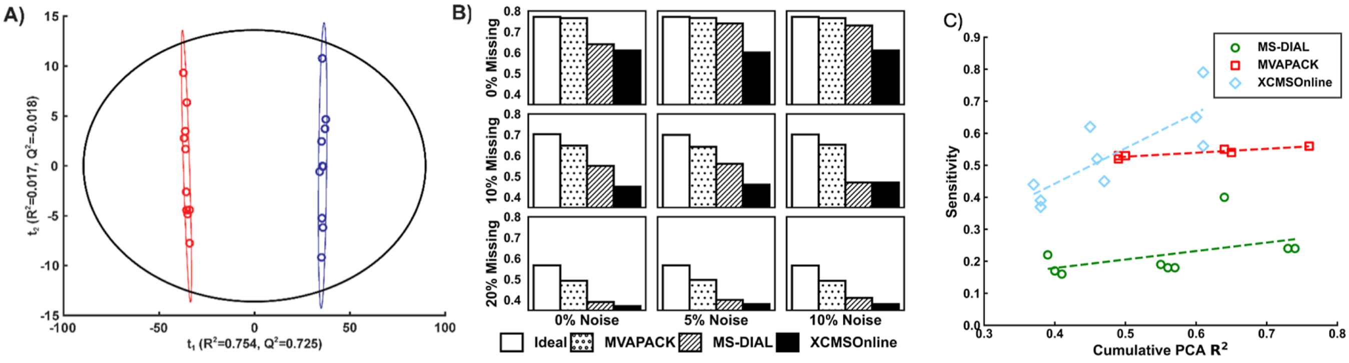


Figure 3. Comparison of software performances based on PCA models created from the simulated LC−MS metabolomics data set. (A) An example of a PCA scores plot generated by the new LC−MS-based functions implemented into the MVAPACK toolkit. The R^2 and Q^2 scores are 0.754 and 0.725, respectively. (B) A compilation of the cumulative R^2 scores for the first two components of the PCA models produced by MVAPACK, MS-DIAL, and XCMSOnline as a function of increased noise and the percentage of missing features. The first bar in each plot corresponds to the cumulative R^2 scores for the ideal feature matrix representing the ideal result for each data set. Please note that the cumulative R^2 score decreases for the ideal matrix according to the increase in the number of missing features. (C) The cumulative R^2 scores for the first two components of the PCA models produced by MVAPACK, MS-DIAL, and XCMSOnline are plotted versus sensitivity.

provides processing capabilities for LC−MS, 1D ^{1}H NMR, 2D ^{1}H−^{13}C/^{15}N/^{31}P NMR, and multiplatform data sets. The process of generating an LC−MS feature matrix and an example MVAPACK script are detailed in Figure 2A,B, respectively. MVAPACK was used to produce PCA models from the simulated (Figure 3A), standard mixture (Figure 4A), and*M. smegmatis*metabolomics (Figure 5A) LC−MS data sets. The MVAPACK PCA score plots are shown in Figures 4A and 5A compared to plots generated by MS-DIAL (Figures 4B and 5B) and XCMSOnline (Figures 4C and 5C).

Validation of LC−MS Data Processing with Simulated Data Set. The LC−MS functionality added to MVAPACK was benchmarked against XCMSOnline and MS-DIAL using the simulated LC−MS data set with various levels of added noise and missing metabolite features. The missing features corresponded to 0, 10, or 20%. Similarly, the added noise amounted to 0, 5, or 10%. The simulated spectra provided a ground truth corresponding to the 935 known statistically distinct metabolite peak features and an idealized PCA model for comparison. Each software platform performed only the initial feature matrix generation due to potential differences in matrix processing algorithms. The PCA algorithm implemented into MVAPACK was previously shown to agree with the SIMCA-P+ equivalent.[36] The raw feature matrix was independently generated for the nine conditions using the three software platforms and each ground truth matrix. To ensure a fair comparison of software performance, feature normalization, filtration, and PCA model calculations were executed by the same MVAPACK script regardless of the software (see Supporting Information).

Software performance was evaluated by comparing the metabolite features identified by MS-DIAL, MVAPACK, or XCMSOnline to the ground truth. Selected metabolites can be either true positives (TP) when the features exist in the ideal matrix or FP when they are absent. Additionally, a FN is an ideal metabolite feature with no counterpart in the matrix selected by the software platform. The TP, FP, and FN results were then used to calculate sensitivity (Supporting Information eq 3), positive predictive value (PPV) (Supporting Information eq 4), and F1 score (Supporting Information eq 5) values. Overall, higher TP rates, sensitivity and PPV percentages, and F1 scores are desirable, while lower FP and FN rates are desirable. The total number of metabolite features identified, the number of TP, FP, and FN peaks, and the calculated sensitivity, PPV, and F1 score values are listed in Table 3 for each simulated data set and software platform. Initially, we

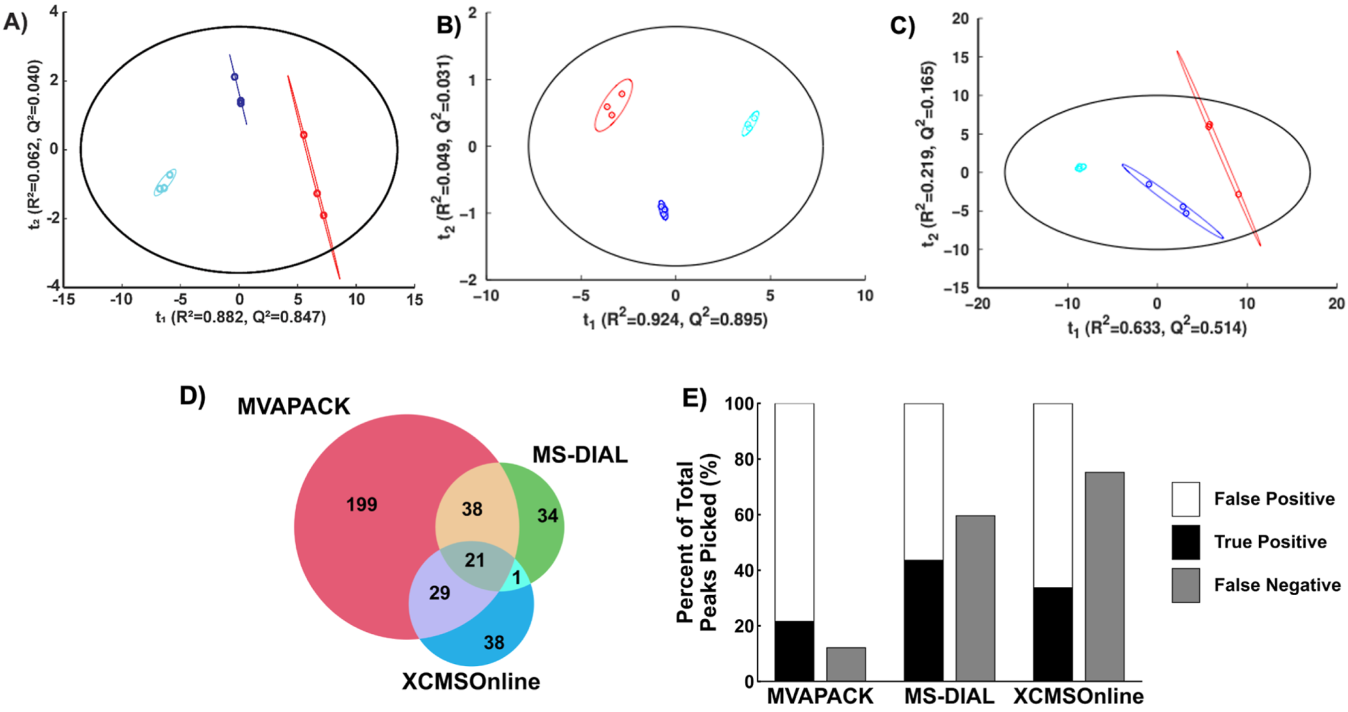


Figure 4. Comparison of software performances based on PCA models created from the standard mixture LC−MS data set. PCA scores plot generated by (A) MVAPACK (R^2 = 0.882, Q^2 = 0.847), (B) MS-DIAL (R^2 = 0.924, Q^2 = 0.895), and (C) XCMSOnline (R^2 = 0.633, Q^2 = 0.514). (D) Venn diagram summarizing the number of features each software platform picked. The identified features were counted as matching between two or more software programs if the m/z and retention times were within 0.1 Da and 10 s, respectively. (E) Bar plot comparing the percentage of TPs, FPs, and FN features to the number of peaks picked by MVAPACK, MS-DIAL, and XCMSOnline.

used the default peak-picking parameters when processing the synthetic data with XCMSOnline and MS-DIAL (Tables S2−S5). However, we observed reduced peak picking performance for both software compared to MVAPACK. We performed a simple grid search for the optimal peak processing parameters of XCMSOnline and MS-DIAL on the 0% missing and noise data set to improve their performance on the synthetic data set. The parameter set with the highest sensitivity was then applied to the other eight versions of the simulated data set (Tables S6 and S7). For XCMSOnline, the ppm peak width and signal-to-noise threshold were optimized to 30 and 12, respectively. For MS-DIAL, the minimum peak height and mass width were optimized to 100,000 counts and 0.05 Da, respectively. We compared the performance of both XCMSOnline and MS-DIAL using their respective optimized parameters for the simulated data set.

Overall, MVAPACK compared well, and in some cases outperformed, both MS-DIAL and XCMSOnline in the analysis of the simulated data sets. None of the software platforms correctly identified all the 935 known and statistically distinct metabolite features in the simulated data set. The best outcome was 735 out of 935 features produced by XCMSOnline when the missing features and added noise were 0%. The worst outcome was 151 out of 935 peaks produced by optimized MS-DIAL parameters at 20% missing peaks and 10% added noise. The average number of TP metabolites detected was 404 ± 164, in which XCMSOnline and MVAPACK were consistently above the average at 498 ± 127 and 509 ± 15, respectively. MS-DIAL was consistently below the average at 206 ± 69. Of course, the outcomes for the FN features were simply the inverse of the TPs. The average number of FNs was 531 ± 164, with both XCMSOnline and MVAPACK consistently below this average at 437 ± 127 and 426 ± 15, respectively. Again, MS-DIAL was consistently above the average at 729 ± 69.

The average number of FP features detected was 231 ± 128, in which XCMSOnline and MVAPACK were consistently below the average at 139 ± 121 and 173 ± 9, respectively. MS-DIAL was significantly above the average at 380 ± 120. Thus, XCMSOnline and MVAPACK performed similarly, having overall stronger performances than MS-DIAL in FP, TP, and FN counts. Despite a lack of optimization, MVAPACK generally performed above the average, slightly better than XCMSOnline. MVAPACK's performance remained consistent despite any noise increase or missing metabolite features. This is evident by the relatively small standard deviations observed for all the values summarized in Table 3. Under all these conditions, the performance of MS-DIAL and XCMSOnline deteriorated proportional to the decay in the quality of the simulated data.

The difference in MS-DIAL, MVAPACK, and XCMSOnline performance was further evident by comparing the sensitivity and PPV values in Table 3. The average sensitivity and PPV values were 43% ± 18% and 62% ± 20%, respectively. Again, MVAPACK performed better than average in both categories, with an average sensitivity of 54% ± 2% and an average PPV of 74% ± 2%. Conversely, MS-DIAL yielded both a low PPV of 35% ± 2% and a low sensitivity of 22% ± 7%, and XCMSOnline produced a better-than-average sensitivity of 53% but a significantly higher PPV of 78% ± 4%. Given the large variance between the sensitivity and PPV values for the three software platforms and the difficulty in discerning which parameter is more important, F1 scores (Supporting Information eq 5) were also calculated. The average F1

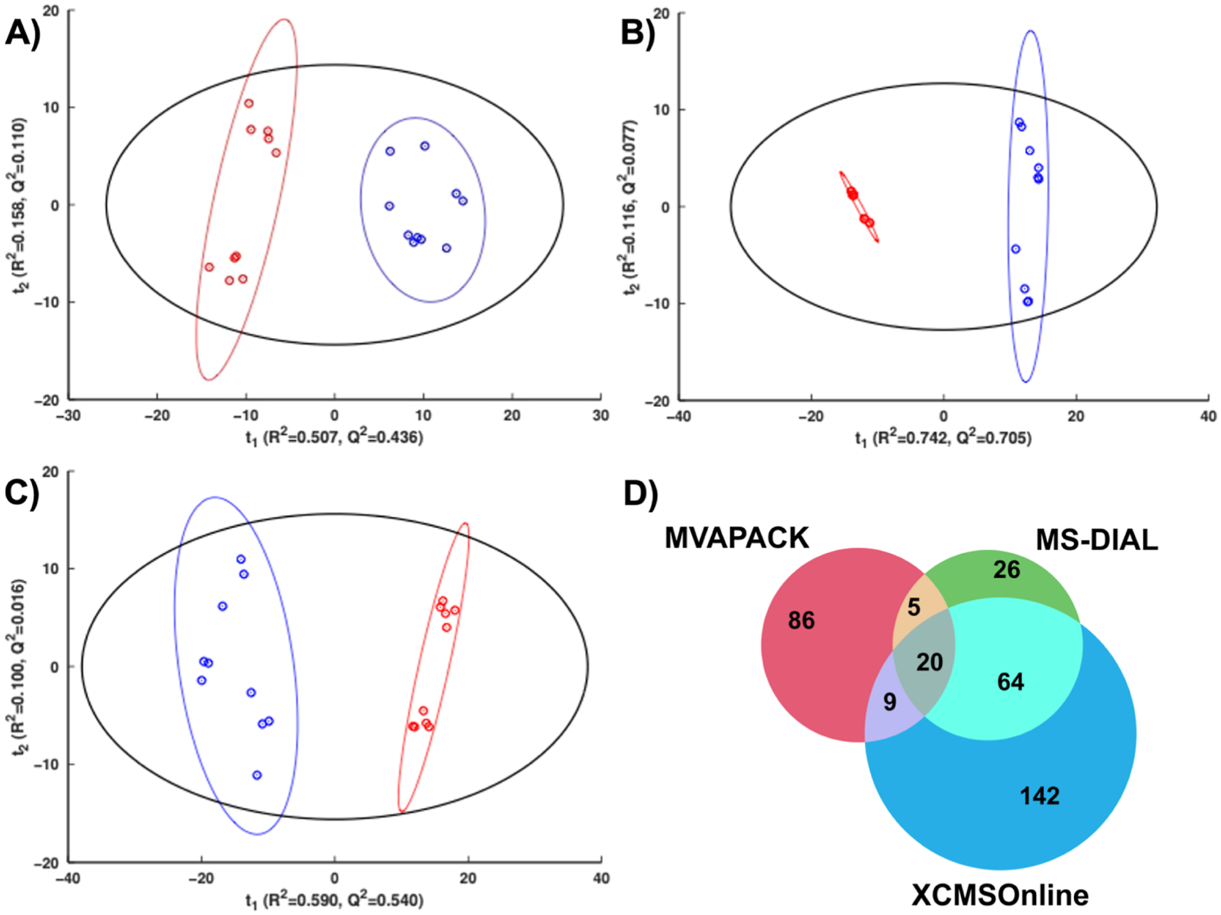


Figure 5. Comparison of software performances based on PCA models created from the*M. smegmatis*LC−MS data set. PCA scores plot generated by (A) MVAPACK (R^2 = 0.507, Q^2 = 0.436), (B) MS-DIAL (R^2 = 0.742, Q^2 = 0.705), and (C) XCMSOnline (R^2 = 0.590, Q^2 = 0.540). (D) Venn diagram summarizing the number of features each software platform picked. The identified features were counted as matching between two or more software programs if the m/z and retention times were within 0.1 Da and 10 s, respectively.

score was 51% ± 19%. MVAPACK had the highest F1 score of 63% ± 2% compared to 27% ± 5% for MS-DIAL and 63% ± 10% for XCMSOnline. This comparison suggests MVAPACK had the best overall performance in the analysis of the simulated data set, although it is rivaled by the performance of XCMSOnline.

Software performance was also assessed by comparing the quality of each PCA model relative to an idealized PCA model calculated from the true matrix of 935 known statistically distinct metabolite features (Figure 3). One method to assess the overall quality of a PCA model is by comparing cumulative R^2 values, where a higher R^2 value indicates a better fit (Table 3). The R^2 values for the first two principal components are plotted in Figure 3B for each data set condition (i.e., different noise levels and percent missing peaks) and the three metabolomics software platforms. Each graph also contains, as the first bar, the idealized results for the true matrix. Under all conditions, MVAPACK performed the best, followed by MS-DIAL and then XCMSOnline. The decrease in performance of MS-DIAL relative to MVAPACK was generally equal to or greater than that of MVAPACK relative to the true matrix. Similarly, the decrease in the performance of XCMSOnline relative to MS-DIAL was often greater than that of MS-DIAL relative to MVAPACK. As expected, the performance of the three metabolomics platforms decreased as the number of missing features and noise levels increased. The performance for the true matrix also decreased as the data set's quality deteriorated. Interestingly, in all cases, the software performance was more sensitive to an increase in the percentage of missing features relative to an increase in noise. The software performance was relatively consistent despite increasing noise at a given percentage of missing peaks (Figure 3B). Similarly, additional false peaks added to the data matrix are added noise that may minimize group differences by diminishing the impact of real peak differences.

A further analysis of the relationship between the cumulative R^2 values and the other performance metrics listed in Table 3 identified significant correlations (Table S8). An average correlation of 0.67 ± 0.17 was observed between sensitivity and the cumulative R^2 values for the three metabolomics software packages (Figure 3C). Sensitivity (Supporting Information eq S3) accounts for TPs and FNs. A similar positive correlation was observed between the cumulative R^2 values and the total number of features identified (0.66 ± 0.15), TPs (0.75 ± 0.22), and F1 scores (0.69 ± 0.14). A negative correlation was observed between missing features (−0.99 ± 0.01) and FNs (−0.75 ± 0.22). Conversely, there was no clear correlation with added noise, PPV, or FPs. Note that PPV (Supporting Information eq 4) accounts for FPs.

The Mahalanobis distance was measured between the two groups in each PCA scores plot and compared to the results obtained from the idealized PCA model (Table 3). A maximal

Table 3. Summary of Performance Metrics for the Simulated Dataset Using Optimized Parameters

software	missing features[a] (%)	noise[b] (%)	real features[c]	features identified[d]	TPs[e]	FPs[f]	FNs[g]	sensitivity[h]	PPV[i]	F1 score[j]	Mahalanobis distance[k]	RV coefficient[l]	R^2[m]
MS-DIAL	0	0	935	1037	374	663	561	40%	36%	38%	1412	0.9988	0.64
MS-DIAL	0	5	935	601	228	373	707	24%	38%	30%	1599	0.9985	0.74
MS-DIAL	0	10	935	693	228	465	707	24%	33%	28%	1950	0.9983	0.73
MS-DIAL	10	0	935	537	180	357	755	19%	34%	24%	757	0.9980	0.55
MS-DIAL	10	5	935	468	166	302	769	18%	35%	24%	666	0.9979	0.56
MS-DIAL	10	10	935	468	166	302	769	18%	35%	24%	543	0.9973	0.57
MS-DIAL	20	0	935	579	205	374	730	22%	35%	27%	105	0.9953	0.39
MS-DIAL	20	5	935	471	156	315	779	17%	33%	22%	170	0.9953	0.40
MS-DIAL	20	10	935	424	151	273	784	16%	36%	22%	157	0.9952	0.41
avg. (STD)				586 (189)	206 (69)	380 (120)	729 (69)	(22) 7	(35) 2	27 (5)	818 (681)	0.9972 (0.015)	0.55 (0.13)
MVAPACK	0	0	935	691	528	163	407	56%	76%	64%	5759	0.9285	0.76
MVAPACK	0	5	935	692	528	164	407	56%	76%	64%	5674	0.9283	0.76
MVAPACK	0	10	935	686	526	160	409	56%	77%	66%	5904	0.9494	0.76
MVAPACK	10	0	935	682	506	176	429	54%	74%	62%	3094	0.8866	0.65
MVAPACK	10	5	935	689	513	176	422	55%	74%	64%	2520	0.8926	0.64
MVAPACK	10	10	935	682	501	181	434	54%	73%	62%	3968	0.9556	0.65
MVAPACK	20	0	935	668	494	174	441	53%	74%	62%	1684	0.9235	0.49
MVAPACK	20	5	935	686	500	186	435	53%	73%	61%	1523	0.8899	0.50
MVAPACK	20	10	935	668	487	181	448	52%	73%	60%	1601	0.9095	0.49
avg. (STD)				683 (9)	509 (15)	173 (9)	426 (15)	54% (2)	74% (2)	63% (2)	3525 (1863)	0.9182(0.025)	0.63 (0.12)
XCMSOnline	0	0	935	889	735	154	200	79%	83%	81%	2567	0.9995	0.61
XCMSOnline	0	5	935	740	612	128	323	65%	83%	73%	1332	0.9994	0.60
XCMSOnline	0	10	935	630	520	110	415	56%	83%	66%	1569	0.9993	0.61
XCMSOnline	10	0	935	758	580	178	355	62%	77%	69%	2217	0.9993	0.45
XCMSOnline	10	5	935	634	489	145	446	52%	77%	62%	1106	0.9992	0.46
XCMSOnline	10	10	935	550	420	130	515	45%	76%	57%	912	0.9991	0.47
XCMSOnline	20	0	935	569	411	158	524	44%	72%	55%	439	0.9990	0.37
XCMSOnline	20	5	935	494	369	125	566	39%	75%	52%	356	0.9988	0.38
XCMSOnline	20	10	935	472	347	125	588	37%	74%	49%	376	0.9988	0.38
avg. (STD)				637 (136)	498 (127)	139 (21)	437 (127)	53% (14)	78% (4)	63% (10)	1208 (754)	0.9991 (0.0002)	0.48 (0.10)
total avg. (Std)[n]				635 (135)	404 (164)	231 (128)	531 (164)	43% (18)	62% (20)	51% (19)	1013 (727)	0.9981 (0.0014)	0.52 (0.12)

[a] Percentage of data metabolite features missing among 10 data files. [b] Percentage of noise synthetically added through waveform functions. [c] Absolute number of real features present in the data set. [d] Number of features identified by specified software. [e] Number of true features identified by the software. [f] Number of false features identified by the software. [g] Number of real metabolite features not identified by the software. [h] Sensitivity as defined by Supporting Information eq 3. [i] PPV of the software as defined by Supporting Information eq 4. [j] F1 score of the software as defined by Supporting Information eq 5. [k] Mahalanobis distance between the two groups in the PCA scores plots. [l] RV coefficient[42] calculated between the PCA scores from the idealized PCA model and each individual PCA model calculated for the simulated data sets. [m] Cumulative R^2 values calculated for each PCA model. [n] Averages and standard deviations include unoptimized MVAPACK values from Table 3.

Table 4. Summary of Performance Metrics for the Standard Mixture Dataset Using Optimized Parameters

software	real features[a]	features identified[b]	TPs[c]	FPs[d]	FNs[e]	sensitivity[f] (%)	PPV[g] (%)	F1 score[h] (%)	Mahalanobis distance (G1−G2)[i]	Mahalanobis distance (G1−G3)[j]	Mahalanobis distance (G2−G3)[k]	RV coefficient[l]	R^2[m]
MS-DIAL	97	94	41	53	56	42	44	43	553	874	712	0.9340	0.97
MVAPACK	97	287	62	225	35	64	22	32	277	660	519	0.7527	0.94
XCMSOnline	97	89	30	59	67	31	34	32	1662	5080	211	0.9900	0.85
average		157	44	112	53	46	33	36	831	2205	481	0.892	0.92
standard deviation		112	16	98	16	17	11	6	733	2492	253	0.124	0.06

[a] Number of real peaks manually identified in the LC−MS spectra. [b] Number of peaks identified by each software. [c] Number of peaks correctly identified by each software. [d] Number of additional peaks identified by the software not assigned as real peaks in the LC−MS spectra. [e] Number of peaks not identified by the software that were present in the spectra. [f] Sensitivity as defined by Supporting Information eq 3. [g] PPV of the software as defined by Supporting Information eq 4. [h] F1 score of the software as defined by Supporting Information eq 5. [i] Mahalanobis distance between the first two groups of PCA scores. [j] Mahalanobis distance between the first and third groups of PCA scores. [k] Mahalanobis distance between the second and third groups of PCA scores. [l] RV coefficient[42] calculated between the PCA scores from manually determined matrix and the automated processes. [m] Cumulative R^2 values calculated for each PCA model.

group separation with minimal within-group variation would result in a large Mahalanobis distance where the idealized PCA model yielded the largest possible Mahalanobis distance for the simulated data sets. We previously demonstrated the utility of the Mahalanobis distance for benchmarking PCA quality and validity.[38] Furthermore, we have identified that the inclusion and amount of noise in the data matrix are important contributors to the quality of a multivariate statistical model.[38−41] MVAPACK produced the largest group separation in the PCA score plots with an average Mahalanobis distance of 3525 ± 1863 and a maximum distance of 5904, indicating the best overall group separation. The average group separation in the MVAPACK PCA model was nearly three times as great as both MS-DIAL and XCMSOnline. XCMSOnline performed slightly better than MS-DIAL with an average and maximum Mahalanobis distance of 1208 ± 754 and 2567, respectively, compared to 818 ± 681 and 1950. These differences are not significant. Not surprisingly, the Mahalanobis distance decreased as more noise was added, TP features were removed, and more FN peaks were detected. Like other statistical parameters, the number of missing metabolite features had the greatest detrimental effect on group separation in the PCA scores plots. The increase in missing features was typically correlated with a decrease in TPs and an increase in FNs. Again, MVAPACK performed better than MS-DIAL and XCMSOnline, maintaining a larger group separation despite the deterioration in the data set's quality.

The idealized PCA models represent the best possible outcomes from the analysis of the simulated data set. In this regard, the quality of the MS-DIAL, MVAPACK, and XCMSOnline PCA models can be further assessed by comparing the scores to this idealized PCA model. The similarities in the PCA scores were determined by calculating the RV coefficients[42] and the squared Pearson correlation. Accordingly, the RV coefficient ranges from 0 to 1, where a score of 1 indicates the two sets of scores are identical. When using default parameters, MVAPACK had the highest average RV coefficient of 0.92 ± 0.03 and MS-DIAL and XCMSOnline had slightly lower and similar results with average RV coefficients of 0.86 ± 0.03 and 0.83 ± 0.04, respectively (Table S5). This trend was inverted with optimized parameters, where MS-DIAL and XCMSOnline had much higher RV coefficients of 0.9972 ± 0.015 and 0.9991 ± 0.0002, respectively. Surprisingly, unlike the other statistical parameters, the RV coefficient did not vary as a function of added noise or missing metabolite features.

Validation of LC−MS Data Processing with the Standard Mixture Data Set. All three metabolomics software platforms were used to independently analyze an LC−MS data set derived from a nine-compound standard mixture (Table 2). The identical protocol was followed as described above for the simulated LC−MS data set. There were nine mzML files comprising three groups, each containing three replicates. The three groups differed by a simple dilution factor. The LC−MS data set was interrogated manually to confirm 97 real spectral features, including eight of the nine standards (Table 4). Reserpine was not identified manually or by any software. Due to poor initial performance with default values, a minimum intensity filter of 1×10^6 counts was added to the peak picking protocols for the three programs. The total number of metabolite features identified, the number of TP, FP, and FN peaks, and the calculated sensitivity, PPV, and F1 score values are listed in Table 4. Analysis was initially done using default parameters for each package, but the resulting metrics demonstrated a need for further refinement (Table S9). Further XCMSOnline and MS-DIAL parameter optimization was performed through a grid search of various parameters on the three replicates in the QC group. Picked peaks were compared to the manually curated list of 97 features. The parameter set with the highest sensitivity was then applied to the other two groups in the standard data set (Tables S10 and S11). All tables and figures in the main text use these optimized parameters for XCMSOnline and MS-DIAL.

Surprisingly, MVAPACK, MS-DIAL, and XCMSOnline performed worse in the analysis of the standard mixture data set compared to the simulated data sets. The decrease in performance may be attributed to fundamental differences between simulated and real experimental data sets. The presence of batch drift, spectral artifacts, chemical noise, larger variance and dynamic range in peak intensities, and nonuniform noise all contribute to a more challenging analysis of an experimental data set that is difficult to reproduce with a simulated spectrum. Overall, MVAPACK, MS-DIAL, and XCMSOnline identified 287, 94, and 89 peaks, respectively (Table 4, Figure 4D). Of these 470 peaks, only 21 were shared by all three software, while 68 peaks were identified by two different programs.

None of the software platforms correctly identified all 97 of the manually verified spectral features associated with the nine-compound standard mixture (Table 4, Figure 4E). The best outcome was 62 out of 97 features produced by MVAPACK. MS-DIAL was second with the correct identification of 41 features. XCMSOnline was the worst performer, with only 30 identified features out of 97. This is somewhat consistent with the performance observed with the simulated LC−MS data set in which MVAPACK showed the strongest performance, although MS-DIAL now slightly edges out XCMSOnline. The average number of TP metabolites detected was 44 ± 16. Of course, the outcomes for the FN features were simply the inverse of the TPs. The average number of FNs was 53 ± 16, with XCMSOnline above this average at 67.

Regarding FPs, the trend was again the reverse of the results obtained with the simulated data sets. MS-DIAL identified the fewest FPs, with the best outcome of 53 FP features. The worst outcome was 225 FPs produced by MVAPACK, with XCMSOnline second with 59 FPs. The average number of FPs detected was 112 ± 98. Thus, both MS-DIAL and XCMSOnline sacrificed identifying TPs to reduce the number of FPs. Conversely, MVAPACK prioritized the identification of TP peaks at the expense of a higher FP rate.

The difference in the performance of MS-DIAL, MVAPACK, and XCMSOnline in the analysis of the standard mixture data set was further evident by comparing the sensitivity and PPV values in Table 4. The average sensitivity and PPV values were 46% ± 17% and 33% ± 12%, respectively. The standard mixture analysis obtained PPV values of 34, 44, and 22% for XCMSOnline, MS-DIAL, and MVAPACK. This disparity in scoring also extends to recalls, where XCMSOnline, MS-DIAL, and MVAPACK have sensitivities of 31, 42, and 64%, respectively. Notably, the average recalls, 43% ± 18% and 46% ± 17%, obtained for the simulated and standard mixture data sets were statistically equivalent. Conversely, the PPV of 33% ± 11% for the standard mixture data set was notably worse and statistically different than the 62% ± 20% obtained with the simulated data set. F1 scores were again calculated further to clarify performance differences between the three software platforms. The average F1 score was 36% ± 6%, again statistically worse than the 51% ± 19% obtained with the simulated data set. MS-DIAL had the highest F1 score of 43% compared to 32% for MVAPACK and XCMSOnline. This comparison somewhat flips the order of MS-DIAL and MVAPACK compared to the outcome with the simulated data sets.

Software performance was also assessed by comparing the quality of each PCA model relative to an idealized PCA model calculated from the true matrix of 97 known spectral features. A simple visual inspection of the PCA scores plots shows a clear separation between the three groups (Figure 4A−C). The overall appearance of the MVAPACK and MS-DIAL PCA scores plots are similar, consistent with the overall similarity in global feature identification. Conversely, XCMSOnline is noticeably different, primarily due to the limited displacement along PC2 between the three groups. Also, the magnitude of the PC1 variance (20) is somewhat higher when compared to MVAPACK (15) and MS-DIAL (10). Again, this is consistent with the difference in the performance of MS-DIAL in its feature identification. To further quantify group separations, the Mahalanobis distance was measured between the three groups in each PCA scores plot and compared to the results obtained from the idealized PCA model. The XCMSOnline PCA model produced the largest Mahalanobis distances, ranging from 211 to 5080. MVAPACK was second, with distances ranging from 277 to 660, close to the averages for the three software platforms. Again, the MS-DIAL and MVAPACK ranking was flipped relative to the simulated data sets. XCMSOnline exhibited the smallest distances, an order of magnitude smaller than MS-DIAL and consistent with the limited number of identified features. PCA score plots were also assessed by calculating the RV coefficients relative to the idealized PCA model. XCMSOnline had the highest RV coefficient of 0.9900 and MS-DIAL was a close second with an RV coefficient of 0.9340. MVAPACK had the lowest RV coefficient of 0.7527.

Validation of LC−MS Data Processing with Biological Data Set. To further assess software performance, we used a realistic biological data set derived from*M. smegmatis*cell lysates with or without treatment with D-cycloserine (DCS), a second-line drug for tuberculosis. The data set consists of two groups, controls versus treatment, each comprising ten biological replicates for 20 LC−MS spectra. We previously reported the major metabolic changes a DCS treatment induced into the*M. smegmatis*metabolome using NMR.[43,44] However, this information cannot be directly correlated to an LC−MS data set. Unlike the two synthetic data sets, this biological data set has an unknown ground truth since the data set was too complex for manual analysis to achieve a reliable ground truth. Accordingly, it was not possible to determine the number of TP, FP, and FN peaks or to calculate sensitivity, PPV, F1 score, and RV coefficients. Instead, we can only provide a limited assessment of the quality of PCA models and the similarities of the data matrices for each software platform (Figure 5). All analyses of the*M. smegmatis* data set utilized the XCMSOnline and MS-DIAL parameters optimized for the standard data set described in the previous section.

The three PCA scores plots appear similar, exhibiting good group separation and comparable within-group variance (Figure 5A−C). A closer examination does highlight a few differences. While the PC2 separation is nearly identical for all packages (−20, 20), PC1 separation is similar for MS-DIAL and XCMSOnline (−40, 40) but less for MVAPACK (−30, 30). The differences in the R^2 and Q^2 quality factors were more relevant. The R^2 values were 66.5, 85.8, and 69.0%, and the Q^2 values were 54.6, 78.2, and 55.6% for MVAPACK, MS-DIAL, and XCMSOnline, respectively. Again, MVAPACK and XCMSOnline performed similarly and worse than MS-DIAL. The selected features show a further divergence between the MVAPACK, MS-DIAL, and XCMSOnline data matrix. A total of 120, 115, and 235 features (Figure 5D) were identified by MVAPACK, MS-DIAL, and XCMSOnline, respectively. A set of 20 features were identified by all three software platforms, and another 78 features were commonly selected by two programs. XCMSOnline selected the most peaks, of which 60.4% had no match with a data matrix from the other two programs. A similar outcome was seen with the simulated LC−MS data set. MS-DIAL and MVAPACK had lower and comparable relative exclusion feature counts with 22.6 and 71.7% of their peaks, respectively, not identified by another software package.

■ DISCUSSION

A novel LC−MS metabolomics data set comprising simulated and experimental data was presented as a valuable and important asset for systematically benchmarking new functionality implemented into our MVAPACK toolkit.[36] The LC−MS data set also serves as a performance comparison tool, which allowed us to evaluate the capabilities and reliability of MVAPACK relative to other software packages routinely used by the metabolomics community. While new LC−MS packages are frequently evaluated with experimental data, a true validation of performance and an assessment of accuracy and reproducibility is difficult or impossible to achieve without ground truth. Furthermore, a negative or biased performance outcome may result from a manually curated LC−MS data set. The inherent complexity of the data may easily lead to the inclusion of any number of FP and FN features identified as true. In this context, a new algorithm or software could produce a correct result and a better analysis of the data set but, nevertheless, be heavily penalized concerning existing software due to these annotation errors that are incorrectly perceived as being correct. Simply put, the algorithm may accurately identify these features as FNs or FPs, but the evaluation metrics would score these as incorrect classifications, downgrading the overall performance of the software. Instead, a simulated data set provides full control over the construction, structure, and content of the data set to establish a ground truth. The user has full control over the number of biological replicates per group, the number of groups, the spectral signal-to-noise, the present or absent peaks, and relative peak intensities and peak shapes. Furthermore, any statistical significance between peaks, spectra, or groups can be easily specified in the design of a simulated data set. An experimental data set derived from a standard mixture may also provide a similar ground truth, but, importantly, under real spectral conditions. A lower level of spectral complexity than a simulated data set is achieved due to a limited number of metabolites in the standard mixtures and a corresponding decrease in spectral features. Further, not all the tunable parameters in simulated spectra can be similarly adjusted in an experimental spectrum of a known mixture. Thus, a simulated and experimental data set can provide useful, complementary information when assessing software performance.

Knowing which peaks and metabolite features exist within the simulated or experimental spectra allows for accurately calculating TP, FP, and FN rates. Peak heights can also be easily defined and varied across replicates to establish statistically distinct features that are known to differentiate between the defined groups. A large, nonstatistically valid variance in peak heights can be assigned to other spectral features to establish a known complex background and complicate the search for differential features. The group-independent variance in random features can be seen as surrogates for chemical noise, impurities, or simply metabolites that do not respond to the external stressor or genetic mutation. Further, the precise control over noise levels and missing spectral features provides a mechanism to stress-test any algorithm as a function of spectral quality. This is easily accomplished with a high level of control in a simulated data set by simply changing peak intensity values. Still, an experimental data set of a standard mixture requires careful and precise adjustments of metabolite concentrations. Accordingly, our simulated LC−MS metabolomics data set consists of 2673 metabolite features, of which 935 are statistically distinct between the two defined groups, comprising 10 replicates each. The 935 statistically significant features have an FC greater than 2 and a CV less than or equal to 25%.[19−22] Our data set consists of 9 sets of spectra (Table 1), where the quality of each subsequent set of spectra is decreased by the addition of an increasing amount of noise (0, 5, 10%) and missing spectral features (0, 10, 20%). Creating a simulated data set with nine distinct quality levels enables granular analysis of algorithm sensitivity to input replicate quality. Likewise, the small size of the files (approximately 50 megabytes per mzML file) enabled rapid analysis of the files by each algorithm used. In addition to the simulated data set, our LC−MS data set for assessing software performance includes an experimental LC−MS data set consisting of a standard mixture of 9 metabolites (Table 2) that yielded 97 annotated features from the manual analysis. The data set contained three groups of ten replicates where the three groups were differentiated by serial dilutions, 1:1, 1:2, and 1:4. A second LC−MS experimental data set is also included that consisted of a polar extract of lysed*M. smegmatis*cells, a nonpathogenic surrogate for tuberculosis. The data set contained two groups of ten replicates where one set of cell cultures was treated with a sublethal dose of DCS, a second-line treatment of TB.[45] Our simulated and experimental LC−MS metabolomics data sets are freely accessible to benchmark the performance of new or existing software or algorithms (https://git.unl.edu/powers-group).

The initial implementation of our MVAPACK software provided a complete processing pipeline from raw 1D or 2D NMR metabolomics data to validated statistical models.[36] Accordingly, MVAPACK provided a diversity of existing functions for the inputting, preprocessing (alignment, normalization, scaling, denoising, etc.), statistical modeling (PCA, OPLS, LDA, etc.), visualization (scores plots, backscaled-loadings plots, S-plots, volcano plots, etc.), and validation (CV-ANOVA, cross-permutation, tree diagrams, etc.) of a data matrix. Thus, incorporating the complete analysis and processing of GC/LC−MS metabolomics data into MVAPACK only required adding new functions to convert standard GC/LC−MS metabolomics data files (e.g., mzML, mzML, proteoWizard.txt) into a data matrix compatible with existing MVAPACK functions. This was achieved by adding 40 new functions written in GNU Octave to the current distribution of MVAPACK (Table S1). Accordingly, MVAPACK now provides a complete data processing pipeline for LC−MS metabolomics data consisting of input/output, peak alignment, feature identification, normalization, imputation, and general data processing. It is important to note that the functions added to MVAPACK were all previously described in the scientific literature and correspond to established and widely used algorithms. No new algorithms were developed for this implementation of the LC−MS processing pipeline. To our knowledge, MVAPACK is now the only metabolomics toolkit capable of providing a complete data processing pipeline for GC/LC−MS, 1D, and 2D NMR metabolomics data sets. MVAPACK can also mix multiple data sets from distinct analytical sources to create a unified statistical model by using multiblock versions of PCA, PLS, and OPLS methods.[46,47] Figures 3A, 4A, and 5A provide representative PCA score plots calculated by MVAPACK utilizing the simulated, standard mixture, and*M. smegmatis*LC−MS metabolomics data sets as diagramed in Figure 2.

MVAPACK users can rely on the package as an all-in-one toolbox for building high-level statistical models from raw spectra. The diversity of supported input data types is a major asset for MVAPACK as researchers have significantly lower technical overheads when conducting metabolomics data analyses. Analysis of MVAPACK, XCMSOnline, and MS-DIAL parameters (Tables S2−S4) show that MVAPACK's default settings are consistent with those in other packages (Table S12), indicating it is similarly suited to general-purpose use. LC−MS analysis with MVAPACK can be performed in less than an hour, regardless of input data size. Accessibility to the underlying functions allows users to work up mzML files in parallel as peak picking relies only on data in a single file. Dozens of input files can be handled in parallel across batch jobs on UNIX-based clusters, which are commonplace in academic research environments. Data sets containing tens of thousands of peaks, metabolites feature matrix refinement, and selection are performed in minutes. This is in stark contrast to desktop and Web server implementations of LC−MS metabolomics processing, which analyze data in serial, leading to analyses taking hours or upward of a day or more to complete. MVAPACK can also be run on a single processor but still provides versatility for researchers.

An exhaustive benchmarking of MVAPACK was undertaken to ensure our software's accurate and reliable performance by comparing its output with that of other popular software in the field. MS-DIAL and XCMSOnline were selected as standards, given their prominence in the metabolomics community and free accessibility. The performance of the three metabolomics software platforms was compared using our LC−MS simulated data set and two experimental data sets. These three LC−MS data sets provided performance benchmarking across various data quality levels. The simulated data set is small, relatively clean, and provides a known ground truth for assessing peak picking performance and metabolite feature identification. The simulated data set has increased levels of real noise and missing peaks but still represents a relatively simple composition to analyze compared to true experimental spectra. The LC−MS spectra collected with real biological samples contain chemical noise, spectral noise, spectral artifacts, and batch-order drift commonly encountered with experimental metabolomics data sets. Accordingly, the experimental data sets provide a more severe stress test for the software packages, but at the expense of limited or no-known ground truths. Overall software performance was quantified and compared by measuring sensitivity, PPV, and F1 scores calculated from the TPs, FPs, and FNs. PCA models were also collected from the data matrix produced by the three software platforms and compared to a PCA model created from an ideal matrix. The ideal data matrix could only be defined from the simulated and standard mixture data sets.

Our benchmarking (Tables 3, 4, and Figure 3B) demonstrates that MVAPACK has a similar or better performance relative to XCMSOnline and MS-DIAL. MVAPACK's performance was less sensitive to increased noise or metabolite feature removal than the other two packages. Not surprisingly given their similar peak identification algorithms, MVAPACK's overall performance was closest to XCMSOnline, where the relative performance ranking was somewhat reversed between the simulated and standard mixture data sets. Conversely, MS-DIAL exhibited notably worse performance on the synthetic data set but improved significantly on the standard and biological data sets. Again, this grouping in performance can be partially explained by the fact that the peak picking used by MVAPACK was based on the Gaussian derivative waveform used by XCMSOnline. XCMSOnline identified significantly more features and fewer FPs than MVAPACK and MS-DIAL with the simulated LC−MS data set. However, this performance decreased considerably as the spectral quality was reduced, a trend reversed for the standard mixture data set. All three software programs performed significantly worse with the standard mixture data set than the simulated data set in all facets except PCA model quality. The simulated data set demonstrated that a decrease in the overall quality of the LC−MS data set based on an increase in noise and missing peaks lead to a pronounced decrease in software performance. Thus, the decrease in software performance with the standard mixture data set represents a continuation of this trend. The level of noise and artifacts was higher in the experimental data sets relative to the simulated data set. The software performance factors obtained with the lowest quality simulated data set (10% noise, 20% missing peaks) were on par with the results obtained with the standard mixture data set. Differences in performance are difficult to explain through parameter comparison alone as all three packages share few analogous parameters, and those shared values are generally comparable and were optimized when different (Tables S2−S4, S6−S8, S10−12). Our results show the software packages may be marginally suited to different LC−MS analysis tasks as each package exposes a mostly unique set of parameters that cannot be readily accessed in comparable packages. It should also be noted that results for any software platform may be improved via parameter tuning, and our results suggest that parameter tuning can largely remove differences, though some will still exist.

In general, the PCA models created by all three software platforms for either of the three data sets were quite visually comparable despite clear differences in the data matrices used to create the multivariate statistical models (Figures 4A−C and 5A−C). The resulting PCA scores plots showed excellent group separation and reliable models. The cumulative R^2 values in Figure 3B provided the best comparative evaluation of software performance based on the PCA model. MVAPACK performed notably better than MS-DIAL and XCMSOnline and was comparable to the ideal model. PCA models generated by XCMSOnline had the lowest cumulative R^2 across all spectral quality settings. Notably, the cumulative R^2 values were shown to correlate with most of the performance metrics listed in Table 3, indicating this composite metric robustly described the overall quality of the data matrix created by MVAPACK, MS-DIAL, and XCMSOnline (Table S8). The cumulative R^2 values nicely capture the overall decay in software performance as a function of missing features and added noise, which leads to the observed increase in FPs and FNs peaks and a decrease in TP peaks. An interesting observation was the significantly higher sensitivity by all three software platforms to missing features relative to increased added noise to the data matrix. The cumulative R^2 values exhibited no correlation with the amount of added Gaussian noise, while the correlation with missing features closely approached the unit in magnitude (−0.98 ± 0.01). Nevertheless, missing true features or additional false features can be considered as alternative sources of noise added to the data matrix. It may be viewed as an extreme version of added Gaussian noise (i.e., mask true peaks).

Quantifying differences between the PCA models by other means, such as measuring Mahalanobis distances between groups, was not particularly informative given similar large group separations. The only notable exception was observed with the XCMSOnline analysis of the standard mixture data set. The relative group separation and PC1 values were significantly higher for XCMSOnline than both MVAPACK and MS-DIAL (Figure 4A−C). Surprisingly, the RV coefficients calculated by comparing the principal components between an ideal PCA model and each calculated model were completely uninformative. It was not platform, noise, or missing feature dependent and seemed to provide random results. Thus, despite the data matrices calculated by each software platform being unique and containing different sets of selected and missed features relative to a perfect data matrix, the RV coefficients were quite similar. An average RV coefficient for the simulated and standard mixture databases were calculated as 0.99 ± 0.00 and 0.89 ± 0.12, respectively. A modest but small difference was observed between the two data sets. Overall, our results suggest a PCA model provides minimal utility in evaluating software performance. This is likely a result of the large, inherent data reduction that is the intended outcome of PCA. The data reduction likely masks intrinsic differences in the original data matrices.

The observation that the PCA model essentially hides underlying differences in the raw data matrices was quite apparent in the performance comparison using the*M. smegmatis*LC−MS metabolomics data set. Visually (Figure 5), the three PCA scores plots were quite similar and yielded the expected outcome of a large group separation between the drug-treated and untreated*M. smegmatis*cell cultures. The addition of DCS, a known antibiotic used as a second line treatment of TB, would be expected to induce a significant metabolic response. Although we observed reasonable overall similarity between the data matrices produced by the three software platforms used to generate the PCA models, the resulting PCA models showed much tighter agreement (Figure 5D). As observed with the simulated LC−MS data set, XCMSOnline and MVAPACK both identified more spectral features than MS-DIAL, with XCMSOnline finding the most overall. Out of the 470 total features identified, a set of 20 were identified by the three metabolomics software platforms, but 78.2, 22.6, and 60.4% of the remaining features were uniquely defined by MS-DIAL, MVAPACK, and XCMSOnline, respectively. In total, 254 spectral features were identified by a single software program. While these software programs and the underlying algorithms are in wide use, these outcomes raise some serious questions about the robustness and reliability of the feature selection protocols. We observed a similar inconsistency in identified lipids when comparing different LC methods for an LC−MS lipidomics study despite all other experimental parameters being identical.[48] Furthermore, a recent article by Li et al. (2023) describing their new Asari LC−MS metabolomics data processing software observed similar discrepancies when comparing its performance to XCMS, MZmine, and MS-DIAL using two experimental LC−MS data sets.[49] Our and other results strongly suggest that further development and understanding of overall performance and reproducibility of metabolomics software is an exciting new direction for research in the community that will stand to benefit the field at large. Our simulated LC−MS data set may provide a valuable asset for these future endeavors.

Herein, we described the creation of a unique metabolomics software package capable of analyzing both NMR and LC−MS data sets. We additionally created a simulated data set for high fidelity benchmarking and performed validation across three distinct data sets as well as a comparison of the results created by MVAPACK versus both MS-DIAL and XCMSOnline. We anticipate that this work will serve as a blueprint for other researchers in the field to perform robust validation of metabolomics packages and algorithms with ours or other similar simulated data sets to advance the field and identify areas of need. While we introduced quality variations, further refinements could be made to optimize the simulated data set by incorporating other commonly encountered LC−MS spectral issues. Future efforts may include adding differences in spectral resolution, peak intensities, and drift times, and the presence of different adducts, oligomers, batch variation, chemical noise, and other spectral artifacts. MVAPACK is freely available for download by interested users at https://bionmr.unl.edu/mvapack.php. Similarly, the corresponding documentation, simulated data set, and all associated scripts can be downloaded from https://git.unl.edu/powers-group/mvapack-lcms-supplemental to both replicate these results and further develop simulated data sets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.3c04979.

Materials and methods related to generation of the simulated data set, the standard mixture, and the standard*M. smegmatis*metabolomics sample; LC−MS data acquisition, software implementation, and validation and benchmarking; and examples of octave/MVAPACK scripts used to process each LC−MS data set. Table S1 lists the LC−MS functions added to MVAPACK. Tables S2−S4 list the MVAPACK, MS-DIAL, and XCMS Online processing parameters; Table S5 lists the performance metrics for the simulated data set using default processing parameters; Tables S6,S7 lists the XCMSOnline and MS-DIal paramaeter optimization for the synthetic data set; Table S8 lists the cumulative R^2 values correlated with performance metrics presented in Table 3; Table S9 lists the performance metrics for the standard mixture dataset using default parameters; Tables S10,S11 lists the XCMSOnline and MS-DIal paramaeter optimization for the standar mixture data set; Table S12 compares the analysis parameters used by MVAPACK, MS-DIAL, and XCMSOnline; Table S13 lists the compound statistics used to generate the simulated LC−MS data set (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Robert Powers − *Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States; Nebraska Center for Integrated Biomolecular Communication, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States*; orcid.org/0000-0001-9948-6837; Phone: (402) 472-3039; Email: rpowers3@unl.edu; Fax: (402) 472-9402

Joseph D. Yesselman − *Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States; Nebraska Center for Integrated Biomolecular Communication, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States*; Phone: (402) 472-2523; Email: jyesselm@unl.edu; Fax: (402) 472-9402

Authors

Christopher P. Jurich − *Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States*

Micah J. Jeppesen − *Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States; Nebraska Center for Integrated Biomolecular Communication, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States*

Isin T. Sakallioglu − *Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States*

Aline De Lima Leite − *Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States; Nebraska Center for Integrated Biomolecular Communication, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.analchem.3c04979

Author Contributions

JDY and RP conceived the project. CPJ contributed new capabilities and functionalities to MVAPACK. ITS, MJJ, and ALL acquired and processed experimental LC−MS data sets. MJJ and ITS validated and tested the MVAPACK software and compared the results with those of other software packages. CPJ wrote the first draft of the manuscript, and all authors contributed to revising it.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under grant no. 1660921 to R.P. and by the National Institute of General Medical Sciences (NIH R35 MIRA 1R35GM147706) to J.D.Y. This work was also supported in part by funding from the Nebraska Center for Integrated Biomolecular Communication (P20GM113126, NIGMS). The research was performed in facilities renovated with support from the National Institutes of Health (RR015468-01).

■ REFERENCES

(1) Idle, J. R.; Gonzalez, F. J. *Cell Metab.* **2007**, *6* (5), 348−351.

(2) Luo, X.; Li, L. *Anal. Chem.* **2017**, *89* (21), 11664−11671.

(3) Ramautar, R.; Mayboroda, O. A.; Somsen, G. W.; de Jong, G. J. *Electrophoresis* **2011**, *32* (1), 52−65.

(4) Huang, Q.; Tan, Y.; Yin, P.; Ye, G.; Gao, P.; Lu, X.; Wang, H.; Xu, G. *Cancer Res.* **2013**, *73* (16), 4992−5002.

(5) Nobeli, I.; Thornton, J. M. *BioEssays* **2006**, *28* (5), 534−545.

(6) Wishart, D. S. *Physiol. Rev.* **2019**, *99* (4), 1819−1875.

(7) Crotty, G. F.; Maciuca, R.; Macklin, E. A.; Wang, J.; Montalban, M.; Davis, S. S.; Alkabsh, J. I.; Bakshi, R.; Chen, X.; Ascherio, A.; et al. *Neurology* **2020**, *95* (24), e3428−e3437.

(8) Arneth, B.; Arneth, R.; Shams, M. *Int. J. Mol. Sci.* **2019**, *20* (10), 2467.

(9) Huang, M.; Zhao, H.; Gao, S.; Liu, Y.; Liu, Y.; Zhang, T.; Cai, X.; Li, Z.; Li, L.; Li, Y.; et al. *Clin. Chim. Acta* **2019**, *497*, 95−103.

(10) Chen, X.; Yu, D. *Metabolomics* **2019**, *15* (2), 22.

(11) Castelli, F. A.; Rosati, G.; Moguet, C.; Fuentes, C.; Marrugo-Ramírez, J.; Lefebvre, T.; Volland, H.; Merkoçi, A.; Simon, S.; Fenaille, F.; et al. *Anal. Bioanal. Chem.* **2022**, *414* (2), 759−789.

(12) Powers, R. *J. Med. Chem.* **2014**, *57* (14), 5860−5870.

(13) Chen, C.-J.; Lee, D.-Y.; Yu, J.; Lin, Y.-N.; Lin, T.-M. *Mass Spectrom. Rev.* **2022**, *42*, 2349−2378.

(14) Kaspy, M. S.; Semnani-Azad, Z.; Malik, V. S.; Jenkins, D. J. A.; Hanley, A. J. *Proteomics* **2022**, *22* (18), 2100388.

(15) Kumar, A.; Misra, B. B. *Proteomics* **2019**, *19* (21−22), 1900042.

(16) Plumb, R. S.; Gethings, L. A.; Rainville, P. D.; Isaac, G.; Trengove, R.; King, A. M.; Wilson, I. D. *TrAC, Trends Anal. Chem.* **2023**, *160*, 116954.

(17) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (12), 1897−1905.

(18) Dettmer, K.; Aronov, P. A.; Hammock, B. D. *Mass Spectrom. Rev.* **2007**, *26* (1), 51−78.

(19) Schiffman, C.; Petrick, L.; Perttula, K.; Yano, Y.; Carlsson, H.; Whitehead, T.; Metayer, C.; Hayes, J.; Rappaport, S.; Dudoit, S. *BMC Bioinf.* **2019**, *20* (1), 334.

(20) Wandy, J.; Davies, V.; Van Der Hooft, J. J. J.; Weidt, S.; Daly, R.; Rogers, S. *Metabolites* **2019**, *9* (10), 219.

(21) Ortmayr, K.; Charwat, V.; Kasper, C.; Hann, S.; Koellensperger, G. *Analyst* **2017**, *142* (1), 80−90.

(22) Xia, J.; Mandal, R.; Sinelnikov, I. V.; Broadhurst, D.; Wishart, D. S. *Nucleic Acids Res.* **2012**, *40* (W1), W127−W133.

(23) Bourgon, R.; Gentleman, R.; Huber, W. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (21), 9546−9551.

(24) Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. *TrAC, Trends Anal. Chem.* **2016**, *82*, 425−442.

(25) Cao, M.; Liu, Y.; Jiang, W.; Meng, X.; Zhang, W.; Chen, W.; Peng, D.; Xing, S. *Sci. Rep.* **2020**, *10* (1), 19524.

(26) Chong, W. P. K.; Thng, S. H.; Hiu, A. P.; Lee, D.-Y.; Chan, E. C. Y.; Ho, Y. S. *Biotechnol. Bioeng.* **2012**, *109* (12), 3103−3111.

(27) Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab. Syst.* **2001**, *58* (2), 109−130.

(28) Bijlsma, S.; Bobeldijk, I.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. K. *Anal. Chem.* **2006**, *78* (2), 567−574.

(29) Yin, P.; Xu, G. *J. Chromatogr. A* **2014**, *1374*, 1−13.

(30) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12* (6), 523−526.

(31) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (3), 779−787.

(32) Jeppesen, M. J.; Powers, R. *Magn. Reson. Chem.* **2023**, *61*, 628−653.

(33) Sansone, S.-A.; Fan, T.; Goodacre, R.; Griffin, J. L.; Hardy, N. W.; Kaddurah-Daouk, R.; Kristal, B. S.; Lindon, J.; Mendes, P.; Morrison, N.; et al. *Nat. Biotechnol.* **2007**, *25* (8), 846−848.

(34) Salek, R. M.; Neumann, S.; Schober, D.; Hummel, J.; Billiau, K.; Kopka, J.; Correa, E.; Reijmers, T.; Rosato, A.; Tenori, L.; et al. *Metabolomics* **2015**, *11* (6), 1587−1597.

(35) Beger, R. D.; Dunn, W. B.; Bandukwala, A.; Bethan, B.; Broadhurst, D.; Clish, C. B.; Dasari, S.; Derr, L.; Evans, A.; Fischer, S.; et al. *Metabolomics* **2019**, *15* (1), 1−5.

(36) Worley, B.; Powers, R. *ACS Chem. Biol.* **2014**, *9* (5), 1138−1144.

(37) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; et al. *BMC Bioinf.* **2008**, *9* (1), 163.

(38) Worley, B.; Powers, R. *Curr. Metabolomics* **2016**, *4* (2), 97−103.

(39) Halouska, S.; Powers, R. *J. Magn. Reson.* **2006**, *178* (1), 88−95.

(40) Vu, T.; Riekeberg, E.; Qiu, Y.; Powers, R. *Metabolomics* **2018**, *14*, 108.

(41) Vu, T.; Siemek, P.; Bhinderwala, F.; Xu, Y.; Powers, R. *J. Proteome Res.* **2019**, *18* (9), 3282−3294.

(42) Robert, P.; Escoufier, Y. *J. Roy. Stat. Soc. C Appl. Stat.* **1976**, *25* (3), 257−265.

(43) Halouska, S.; Chacon, O.; Fenton, R. J.; Zinniel, D. K.; Barletta, R. G.; Powers, R. *J. Proteome Res.* **2007**, *6* (12), 4608−4614.

(44) Halouska, S.; Fenton, R. J.; Zinniel, D. K.; Marshall, D. D.; Barletta, R. G.; Powers, R. *J. Proteome Res.* **2014**, *13* (2), 1065−1076.

(45) Caminero, J. A.; Sotgiu, G.; Zumla, A.; Migliori, G. B. *Lancet Infect. Dis.* **2010**, *10* (9), 621−629.
(46) Xu, Y.; Goodacre, R. *Metabolomics* **2012**, *8* (S1), 37−51.
(47) Worley, B.; Powers, R. *Chemom. Intell. Lab. Syst.* **2015**, *149* (Part B), 33−39.
(48) Sakallioglu, I. T.; Maroli, A. S.; Leite, A. D. L.; Powers, R. *J. Chromatogr. A* **2022**, *1662*, 462739.
(49) Li, S.; Siddiqa, A.; Thapa, M.; Chi, Y.; Zheng, S. *Nat. Commun.* **2023**, *14* (1), 4113.