ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Article

Workflow for Evaluating Normalization Tools for Omics Data Using Supervised and Unsupervised Machine Learning

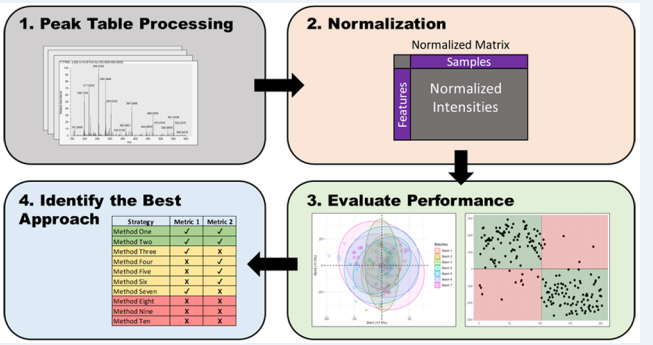
Aleesa E. Chua, Leah D. Pfeifer, Emily R. Sekera, Amanda B. Hummon, and Heather Desaire*

Cite This: *J. Am. Soc. Mass Spectrom.* 2023, 34, 2775−2784

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: To achieve high quality omics results, systematic variability in mass spectrometry (MS) data must be adequately addressed. Effective data normalization is essential for minimizing this variability. The abundance of approaches and the data-dependent nature of normalization have led some researchers to develop open-source academic software for choosing the best approach. While these tools are certainly beneficial to the community, none of them meet all of the needs of all users, particularly users who want to test new strategies that are not available in these products. Herein, we present a simple and straightforward workflow that facilitates the identification of optimal normalization strategies using straightforward evaluation metrics, employing both supervised and unsupervised machine learning. The workflow offers a "DIY" aspect, where the performance of any normalization strategy can be evaluated for any type of MS data. As a demonstration of its utility, we apply this workflow on two distinct datasets, an ESI-MS dataset of extracted lipids from latent fingerprints and a cancer spheroid dataset of metabolites ionized by MALDI-MSI, for which we identified the best-performing normalization strategies.



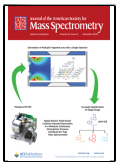
INTRODUCTION

An important step in achieving high-quality omics results is data preprocessing. Data preprocessing aims to enhance the biologically relevant signals, while minimizing the impact of unwanted variation or bias. Some essential preprocessing steps include filtering, peak detection, peak picking, alignment, and normalization. The complexity of the MS data makes the manual processing of raw data tedious. This challenge has spurred the development of processing pipelines like MetaboAnalyst, MZmine 3, and XCMS, which have streamlined the processing of raw MS data.[1−3] The value of these tools is underscored by their widespread popularity, with tools like MetaboAnalyst garnering over 500,000 users globally.[4] While these tools effectively cover most steps of data preprocessing, a key limitation lies in their limited scope in data normalization.

Normalization is critical to the data processing pipeline as it minimizes unwanted systematic or technical variation, which can be introduced to samples via sample preparation and handling or through data acquisition. Normalization is particularly important when biological variation is small as these systematic biases often obscure the more valuable signal variations. Common approaches for normalizing MS data involve linear or global scaling approaches, which involve scaling to a normalization factor.[5] These approaches are commonly incorporated into processing pipeline tools. While XCMS does not provide options for data normalization, MetaboAnalyst and MZmine 3 offer users over six normalization algorithms, most of which encompass these scaling approaches, with strategies like normalizing to the average intensity, a reference feature, or the total raw signal.[1,2]

While global scaling approaches can improve omics analyses, other normalization methods also exist, and choosing the optimal approach is best done empirically. For example, Valikangas and co-workers evaluated the effect of 11 different normalization methods on four different types of proteomic datasets.[6] They identified variance stabilization normalization as the best method for normalizing proteomic data. Another study, by Benedetti and co-workers, evaluated the performance of seven normalization methods on glycan data, for which they identified the combination of probabilistic quotient normalization and a log transformation to be the best-performing approach.[7] These studies highlight the data-dependent nature of normalization; there is no one-size-fits-all approach that is optimal for all types of datasets. The choice of normalization strategy is even further complicated by the many approaches

Received: August 18, 2023
Revised: October 4, 2023
Accepted: October 11, 2023
Published: October 28, 2023



available and the potential for their different combinations. Given the abundance of normalization approaches, how can mass spectrometrists choose the optimal strategy to implement in their dataset?

Several software tools and packages aim to address this issue by providing users with an evaluation of the normalization performance. Tools like NormalyzerDE[8] evaluate strategies using p-value histograms and receiver operating characteristic curves, while tools like NOREVA[9,10] and MetaX[11] implement measures of intragroup variation and an assessment of intensity distributions and PCA scores. Like the processing pipelines, these normalization assessment tools remain limited to the strategies incorporated in them. MetaX[11] only evaluates seven strategies, while NormalyzerDE[8] and NOREVA[9] offer 13 and over 20 strategies, respectively. NOREVA, the most comprehensive of the three, also enables users to evaluate different combinations of these strategies. Use of these tools results in a report, where normalization effectiveness can be assessed using 7+ distinct graphs and plots for each strategy.

These tools may streamline the process of identifying the best approach, yet they have limitations. For example, many users view the task of learning to use a new program, just for the single task of normalizing data, as an undesirable barrier; this barrier is larger for academic software, which is typically not well documented or supported. Furthermore, the tools' requirement that users upload their research data also restricts their usage; many cannot or do not want to share their unpublished data. Finally, both NormalyzerDE and NOREVA output pages of graphs for each normalization strategy, leaving users often scratching their heads, not knowing what metrics, of the very many provided, they should most attend to.

We address the needs of users who prefer to assess normalization methods themselves but need some guidance on how to proceed. We outline a straightforward workflow for the identification of the best-performing normalization strategy for MS datasets, using two key performance evaluation metrics. Specifically, normalization performance is (1) visually assessed by comparing raw and normalized principal components analysis (PCA) plots and (2) quantitatively assessed by comparing supervised classification accuracies and area under the receiver operating curve (AUC) values before and after normalization. By integrating both PCA and supervised classification into the workflow, we provide users with a clear and comprehensive toolbox that captures the unique yet complementary advantages of the two metrics. PCA, as an unsupervised technique, helps identify broad patterns within a dataset. By showing where the most significant variations in the data lie, users can quickly detect issues such as batch effects or better understand the dominant sources of biological variation. This can facilitate the identification of which normalization strategies are better suited for their data. Supervised classification emphasizes the data's ability to partition between specific categories or groups. This is particularly important for studies where prediction accuracy is important, such as biomarker studies. In identifying reliable biomarkers, it is especially important to optimize AUC, as it is generally considered a more useful metric than accuracy, particularly when the classes are imbalanced. By leveraging these two criteria in tandem, users benefit from their complementary relationship: PCA offers a broad understanding of the data, while supervised classification offers granular insights. Together, these metrics offer an easily interpretable understanding of how normalization impacts the data. These chosen metrics not only capture the impact of normalization but are also user-friendly and accessible to all users, regardless of their level of expertise with data science or machine learning.

This workflow is iterative and can accommodate any number of normalization approaches. Furthermore, we provide users with a way to choose the "best" approach; it is identified by comparing the PCA and supervised classification results of all the normalization strategies tested. We demonstrate the utility of this workflow and identify optimal normalization strategies for two datasets: a latent fingerprint dataset of lipids acquired using electrospray ionization mass spectrometry (ESI-MS), and a cancer spheroid dataset containing metabolites ionized by matrix-assisted laser desorption ionization mass spectrometry imaging (MALDI-MSI). While our demonstration focuses on lipidomic and metabolomic data, the workflow can be extended to other omics fields, like proteomics and glycomics, and quantitative MS experiments provided the MS data can be matrix-formatted and batch information is available.

The value of this workflow lies in its simplicity. With a basic understanding of R, the outlined workflow can be implemented and applied to all types of MS datasets for the straightforward evaluation of normalization performance. With this methodology, users can try both old and new normalization strategies, ensuring their adaptability as new research unveils novel ways to normalize MS data. Ultimately, the protocol is easily amenable to scientists' needs, offering a simple workflow to identify the best-performing normalization strategies using only two key metrics.

METHODS

Latent Fingerprint Sample Collection and Preparation. An ESI-MS dataset of latent fingerprints that had been collected for a separate project[12] was used for the studies herein. Briefly, groomed fingerprints were generated by one participant first touching facial regions with a high sebum content (cheek, neck, and forehead) and then depositing a fingerprint onto a piece of aluminum foil. Samples were either prepared right away or allowed to oxidize for 24 h, creating two distinct sample types. Each sample was desalted by liquid−liquid extraction and stored at −20 °C until analysis.

ESI-MS. The MS data of 202 samples were acquired in seven different batches over the course of 1 year using direct infusion on an Orbitrap Fusion Tribrid mass spectrometer (Thermoscientific, San Jose, CA). No two batches were acquired in the same week. Mass spectra were acquired using the negative ion mode, with a spray voltage of 2.3 kV, a resolution of 60k, and a mass range of m/z 150 to 600. The ESI-MS conditions have been previously described.[12] After data acquisition, raw spectral files (.RAW) were converted to .MS1 files using RawConverter (Scripps, Version 1.2.0.0) with the default settings. The data were extracted into a matrix of samples and features of binned intensities using LevR.[12] The following settings for LevR were used: 25% empty cells allowed, number of lines in header is 20, lower m/z of 150, higher m/z of 600, and 0.0125 Da bin width.

Cell Culture and Growth of Spheroids. The human colon carcinoma line HT-29 was obtained from the American Type Culture Collection (ATCC, Manassas, VA). HT-29 cells were grown in McCoy's 5A cell-culture medium (Life Technologies, Grand Island, NY) supplemented with 10% fetal-bovine serum (FBS; Hyclone Laboratories, Logan, UT), 1% L-glutamine (Life Technologies, Grand Island, NY), and 1% penicillin/streptomycin/amphotericin B (Corning, Mana-

https://doi.org/10.1021/jasms.3c00295
J. Am. Soc. Mass Spectrom. 2023, 34, 2775−2784

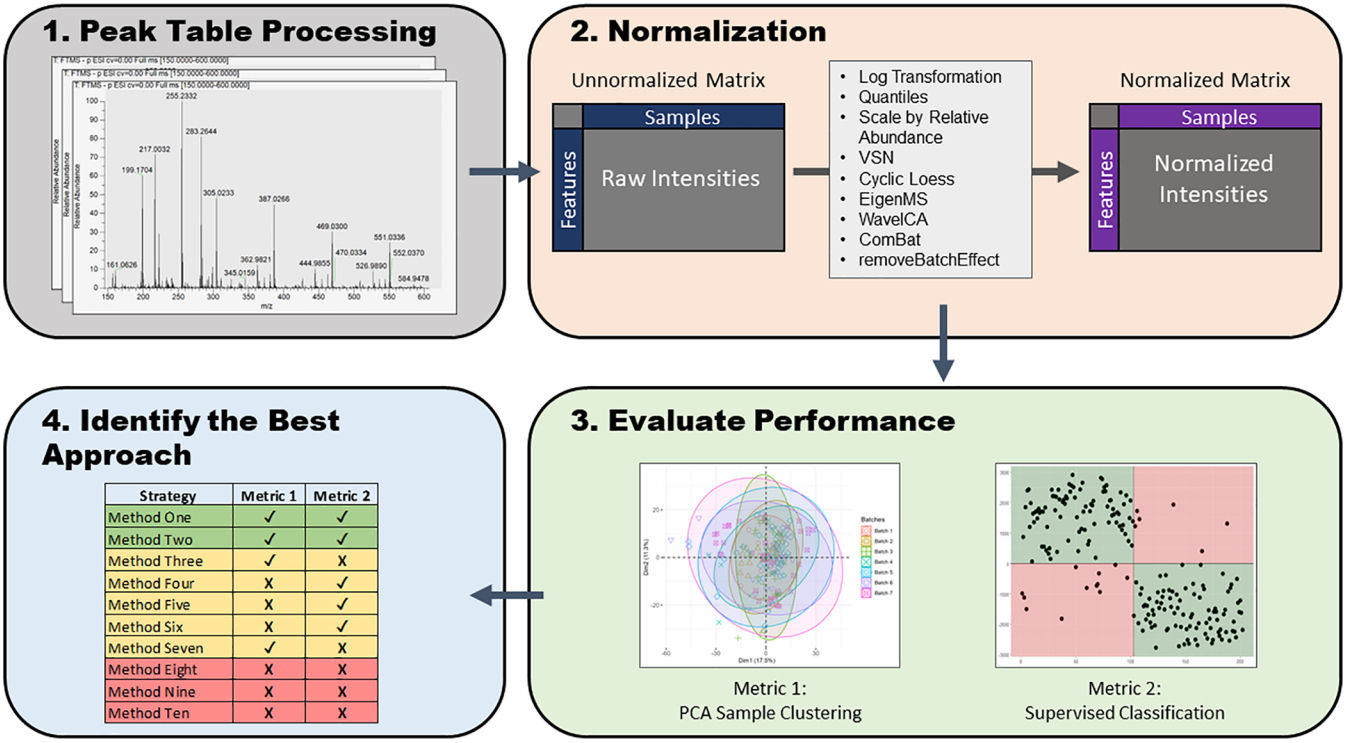


Figure 1. Workflow overview for identifying optimal MS normalization strategies. Raw MS data are processed into a matrix of features and samples, to which normalization strategies can be applied. The resulting normalized matrices are then passed as inputs to unsupervised and supervised classification, where the performance of each approach is evaluated. The metrics of PCA sample clustering and supervised classification outcomes guide the selection of the best approach.

ssas, VA). Mycoplasma testing of the cell line was completed by utilizing a Mycostrip mycoplasma detection kit (Invitrogen, San Diego, CA). Spheroids were prepared in an agarose-coated 96-well plate as previously described.[13−15] Cells were seeded into each well at a density of 7,000 cells per well in 200 μL of media and centrifuged at 1,000*g* for 10 min prior to incubation at 37 °C and 5% CO_2. After an initiation period of 4 days, 50% of the culture volume was replaced with fresh medium every 2 days thereafter. On day 12 of culture, the spheroids reached a diameter of approximately 1 mm, and half of the 96-well plate was harvested for embedding. On day 13, the second half of the spheroids were harvested.

Sample Preparation for MALDI-MSI Analysis. After growth day 12 or 13, medium was aspirated, and spheroids, ten from each day, were washed twice with 1× PBS and embedded following previously published protocols.[16] Briefly, spheroids were transferred into the base of gelatin arrays (20% w/v gelatin), and excess PBS was removed. Warm gelatin was placed on top of the spheroids and flash frozen at −80 °C until sectioning. Spheroids were sectioned at 12 μm thick sections using a cryostat at −30 °C and thaw mounted onto indium tin oxide coated glass slides (Delta Technologies, Loveland, CO). Samples were coated with 9-aminoacridine (9AA) at a concentration of 5 mg/mL in 75% acetonitrile using an HTX Imaging M5 TM-Sprayer (Chapel Hill, NC). Matrix was applied at 60 °C for 8 passes over the samples. The flow rate of matrix was 0.12 mL/min at a velocity of 1,200 mm/min, track spacing of 2 mm, pressure of 10 psi, gas flow rate of 2 L/min, nozzle height of 40 mm, and a drying time of 2 s between each pass. Samples were analyzed immediately after application of the MALDI matrix.

MALDI MSI. MALDI-MSI spectra were acquired using an UltrafleXtreme MALDI-TOF-TOF mass spectrometer (Bruker Daltonics, Bremen, Germany) equipped with a smartbeam II Nd:YAG 355 nm laser. Mass spectra were acquired using the reflectron in negative ion mode with a mass range set to acquire between m/z 200−1,000. The laser spot size was set to "small", at a frequency of 2,000 Hz with the raster distance set to 25 μm along the x and y axes. To obtain optimal signal intensity, each pixel accumulated was set to 500 shots. External calibration was completed using a red phosphorus standard mixture. Images were initially processed with flexImaging 4.1 software (Bruker Daltonics, Bremen, Germany) to convert to a.imzML using a bin size of 32,000 data points.

Data from the 20 different spheroid images were extracted into a single matrix of samples and features following the data extraction protocol we published previously.[17] Each column in the extracted matrix contains data from a single pixel in one spheroid (the samples), and each row represents one of the 32,000 bins of data along the m/z axis, the features. Prior to normalization studies, data were additionally minimally processed, removing pixels whose total ion current (TIC) was less than 200 counts, which corresponds to 1.8% of the maximum TIC for any pixel. This step results in a total of 9156 pixels from 20 different spheroids. Also, the features (bins) were downselected using the following criteria: For each feature, at least 10% of the pixels in the dataset had to have an intensity of greater than 0.12. This threshold represents 0.1% of the overall maximum intensity in the matrix (for any feature or sample.) This feature-reduction step removes bins that are low-abundance and not likely to be useful in classification without introducing bias. After this feature down-selection

https://doi.org/10.1021/jasms.3c00295
J. Am. Soc. Mass Spectrom. 2023, 34, 2775−2784

Table 1. Normalization Strategies Applied to the ESI-MS and MALDI MSI Datasets

Normalization Strategy	Type of Strategy	ESI-MS Dataset	MALDI-MSI Dataset[a]
Log transformation	Transformation to reduce data skewness	+	+
Scale by Relative Abundance	Scaling technique normalizing to ion abundance	+	+
removeBatchEffect	Batch effect removal algorithm	+	+
ComBat	Batch effect removal algorithm	+	+
Relative Abundance + removeBatchEffect	Scaling and batch effect removal	+	+
Relative Abundance + ComBat	Scaling and batch effect removal	+	+
Quantiles Normalization	Nonlinear technique that ensures the distributions across samples have the same quantiles	+	+
Quantiles + removeBatchEffect	Nonlinear normalization and batch effect removal	+	+
Quantiles + ComBat	Nonlinear normalization and batch effect removal	+	+
Cyclic Loess	Nonlinear technique that iteratively applies the loess regression to the MA plot of the data	+	+
Cyclic Loess + removeBatchEffect	Nonlinear normalization and batch effect removal	+	+
VSN	Nonlinear technique that stabilizes the variance across different intensity levels	+	+
EigenMS	Nonlinear technique that combines singular value decomposition and ANOVA to minimize systematic variation	+	–
WaveICA	Batch effect removal using the wavelet transform method with independent component analysis	+	–

[a]EigenMS and WaveICA were not used on the MALDI-MSI dataset due to either an extensive execution time or irrelevance to MALDI data.

Table 2. R Requirements for Normalization

Strategy	R Package/Source	Normalize Function	Input Requirements
Log transformation	-	log2(Mat)	Feature x sample matrix
Scale by Relative Abundance	-	sweep(Mat, 2, scalingfactor_vec, "/")	Feature x sample matrix
removeBatchEffect	BiocManager, limma	removeBatchEffect(logMat, batch = batch_vec)	Input needs to be a log-transformed feature x sample matrix.
ComBat[24]	BiocManager, sva	ComBat(Mat, batch)	Feature x sample matrix
Quantiles	BiocManager, limma	normalizeQuantiles(Mat)	Feature x sample matrix
Cyclic Loess	BiocManager, limma	normalizeCyclicLoess(Mat)	Input needs to be a log-transformed feature x sample matrix.
VSN	BiocManager, limma, vsn	normalizeVSN(Mat)	Feature x sample matrix.
EigenMS[20]	EigenMS.R source code (downloaded through eigenms.sourceforge.net)	>m_ints_eig1 = eig_norm1(m = m_logInts, treatment = grps, prot.info = m_prot.info) > eig_norm2(rv = m_ints_eig1)	Input needs to be a log-transformed feature x sample matrix.
WaveICA[21]	Install using devtools: > devtools::install_github("dengkuistat/WaveICA", host = "https://api.github.com")	WaveICA(Mat, batch)	Input sample x feature matrix needs to be ordered by injection order.

step, the total number of features in the matrix is reduced from 32,000 to 7209.

Data Processing and Analysis. All data analysis was performed in RStudio, R version 4.2.2. Missing values in the MALDI data matrix were imputed by using average intensities. For normalization, the following packages were installed: Biocmanager, limma,[18] and sva.[19] The source code for EigenMS[20] normalization was downloaded from eigenms.sourceforge.net. The WaveICA[21] strategy was installed through github using the devtools package. PCA was performed using factoextra (Version 1.0.7). Supervised classification of unnormalized and normalized matrices were performed using the Aristotle Classifier[22] and XGBoost[23] (R package, xgboost, Version 1.6.0.1). The code for the version of the Aristotle Classifier used can be found in the Supporting Information of ref 22. A classification accuracy score for each strategy was determined for both classifiers by comparing the predictions of the classifier with the actual values. The AUC value of the receiver operating characteristic (ROC) curve for each classifier and DeLong test p-values were determined using the R package, pROC. Example code, along with the two input matrices, can be found in the Supporting Information.

Hyperparameters. For classification by XGBoost, default hyper parameters were used: booster = "gbtree", objective = "binary:logistic", eta = 0.3, gamma = 0, max_depth = 6, min_child_weight = 1, subsample = 1, colsample_bytree = 1. No parameters were optimized to avoid overtraining. The evaluation metric used was "auc", computed by measuring the area under the ROC curve. The XGBoost model was trained using 90% of the data, while the other 10% was used to test the model's performance. For classification using the Aristotle Classifier, the "X" variable, denoting training group size, was set to 6 for both datasets. The length of the classification time increases as the "Repeats" variable (K) increases. Due to the differences in size of the two datasets, two different K values were used: K was set to 1000 for the smaller ESI-MS dataset and 300 for the larger MALDI-MSI dataset.

RESULTS AND DISCUSSION

Overview and Workflow. Reliable analysis of mass-spectrometry-based omics data is dependent on effective data normalization, and the overall goal of this research is to demonstrate a simplified method for identifying the best strategy for normalizing MS data. An overview of the workflow is presented in Figure 1. In this workflow, the raw MS data are converted into a peak table format, or a matrix of samples and features. This matrix can then be applied to several different

https://doi.org/10.1021/jasms.3c00295
J. Am. Soc. Mass Spectrom. 2023, 34, 2775−2784

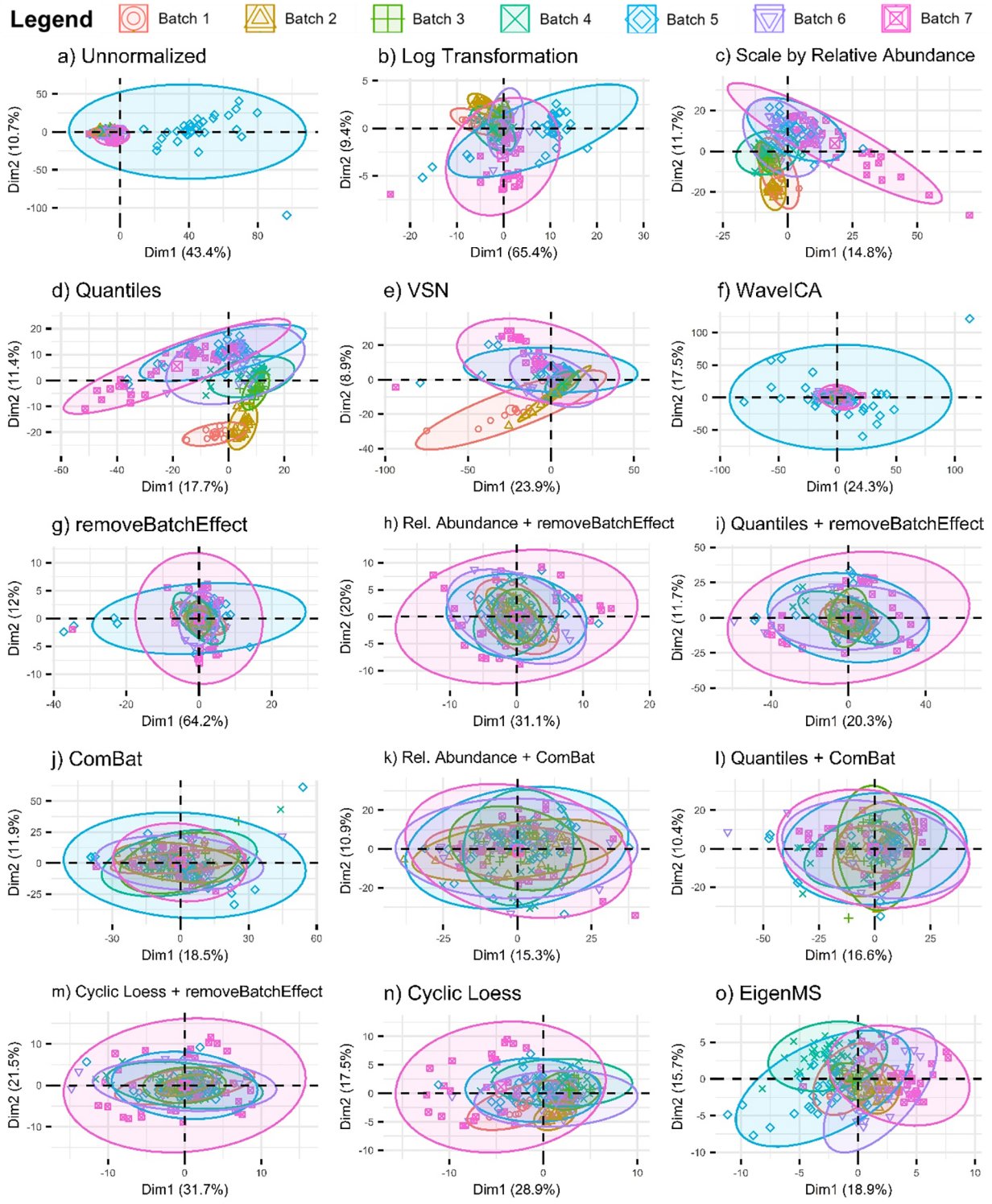


Figure 2. PCA plots for latent fingerprint samples (a) before and (b–o) after normalization. Colors indicate different batches, where batches are defined as the set of samples run on the mass spectrometer in 1 day. For each batch, 95% confidence ellipses are shown.

normalization strategies, where normalization performance is evaluated by comparing the PCA and supervised classification outcomes of the raw and normalized matrices. In this workflow, XGBoost and the Aristotle Classifier were chosen due to their superior performance in the supervised classification of proteomic data.[22] By evaluating the performance of all tested strategies against these two metrics, the best-performing strategy can be identified.

The workflow is applied to two distinct datasets, a latent fingerprint dataset of lipids acquired using ESI-MS and a cancer spheroid dataset of metabolites ionized using MALDI-MSI, to evaluate the performance of 12+ different approaches. These approaches include a log transformation, quantiles normalization,[18] ComBat,[24] removeBatchEffect,[18] scale by relative abundance, EigenMS,[20] WaveICA,[21] CyclicLoess,[18] variance stabilization normalization (VSN),[18] as well as some

https://doi.org/10.1021/jasms.3c00295
J. Am. Soc. Mass Spectrom. 2023, 34, 2775−2784

Table 3. Supervised Classification Outcomes for ESI-MS Dataset[a]

Approach	XGBoost			Aristotle Classifier			
	Accuracy (%)	AUC	DeLong *P*-Value	Accuracy (%)	AUC	DeLong *P*-value	Performance
Unnormalized	82.18	0.9069	-	62.87	0.6455	-	-
removeBatchEffect	88.61	0.9378	0.0392	78.71	0.8426	1.74×10^{-08}	Good
EigenMS	86.14	0.9391	0.0558	88.61	0.9454	1.73×10^{-12}	Good
Cyclic Loess	86.63	0.9363	0.0563	82.67	0.8801	6.94×10^{-08}	Good
Cyclic Loess + removeBatchEffect	85.64	0.9375	0.0622	81.68	0.9094	4.29×10^{-10}	Good
ComBat	82.67	0.9316	0.1084	79.21	0.8399	9.19×10^{-07}	Good
Quantiles + ComBat	84.65	0.9175	0.4980	82.67	0.8954	3.72×10^{-09}	Good
Relative Abundance + removeBatchEffect	85.15	0.9275	0.2117	77.23	0.8433	8.45×10^{-06}	Good
Relative Abundance	79.70	0.8816	0.1574	76.73	0.8103	0.0002	Mediocre
Relative Abundance + ComBat	85.15	0.8968	0.5941	77.72	0.8255	2.99×10^{-05}	Mediocre
Quantiles	82.18	0.8927	0.3987	78.71	0.8533	3.44×10^{-06}	Mediocre
Quantiles + removeBatchEffect	83.66	0.8979	0.6237	81.19	0.8572	2.13×10^{-06}	Mediocre
Log Transformation	82.67	0.9165	0.2040	65.35	0.6417	0.7987	Mediocre
WaveICA	80.69	0.8885	0.5109	51.49	0.4963	0.0039	Poor
VSN	73.27	0.8399	0.0002	54.46	0.5634	0.0079	Poor

[a]Approaches with accuracies and AUC values superior to the unnormalized data for both classifiers are identified as "good" strategies, while approaches that outperform only one classifier are indicated as "mediocre". Approaches displaying lower accuracies and AUC values for both classifiers are marked as "poor".

combinations of these strategies. Batch effect removal algorithms were paired with both scaling and nonlinear techniques, ensuring not to combine two batch-removal algorithms to prevent overfitting. Table 1 shows the normalization strategies applied to each dataset as well as a brief description of the strategies employed.

While any normalization method could be tested using this workflow, we additionally provide researchers with a list of effective strategies, instructions for downloading the R package to apply the strategies, and some expert tips on unique requirements for each one when they are present. See Table 2, which provides nine different strategies that can be applied, the exact function for normalization for each strategy, the data formatting requirements for each normalization method, and the package from where it originates. It is important to note whether the strategy used has specific requirements (such as a log transformation) for the input matrix. In each case, performing the normalize function will generate a normalized matrix.

Normalization of a Latent Fingerprint Dataset of Lipids Acquired Using ESI-MS. *Metric 1: Sample Clustering in Unsupervised Classification.* In PCA, samples cluster based on the greatest source of variance in the dataset. If samples cluster by sample type (such as healthy vs diseased), then the first two principal components likely capture this biological variance. If, however, samples cluster by batch, or the set of samples run on the mass spectrometer in 1 day, then the first two principal components likely capture batch effects. By comparing the PCA plots before and after normalization, the effect of normalization can be visualized. For samples where batch effects are evident, effective strategies will result in PCA plots in which batch effects are minimized; there will be a greater overlap between clusters of different batches. For samples that cluster by sample type, effective strategies result in PCA plots where clusters of different sample types are better separated.

Figure 2 shows the results of unsupervised classification by PCA before and after normalization. In this dataset, the PCA for the unnormalized matrix shows distinct sample clustering by batch (Figure 2a). As batch effects are captured within the first two principal components, the best approaches for this method likely deal with batch variation in their algorithm and will subsequently minimize the percentage of variability captured by these two principal components. The performance of the 14 normalization approaches can thus be categorized based on how well they reduce PCA clustering by batch.

Four methods exacerbate batch effects: log transformation, scale by relative abundance, quantile normalization, and VSN (Figure 2b–e). As samples are more tightly clustered by batch, these are considered poor-performing strategies. For one method, WaveICA, normalization has no impact on alleviating batch effects (Figure 2f). EigenMS, Cyclic Loess, and all algorithms that explicitly deal with batch effects perform well for this metric, resulting in PCA plots in which there is an increased overlap between samples of different batches (Figure 2g–o). These batch effect removal algorithms include removeBatchEffect and ComBat, as well as all their variants.

Metric 2: Supervised Classification Outcomes. An effective normalization strategy would minimize batch effects and enhance the ability to detect the target biological variable, thereby improving the classification accuracy and AUC values. Severe deterioration of supervised classification outcomes observed after normalization may be suggestive of removing "too much" from the data. The normalization method may have removed some biological variation, thus contributing to inferior outcomes. Note that the classification, as performed herein, follows established principals that avoid artificially inflated accuracy.[25]

The performance of the 14 normalization strategies were evaluated and categorized based on differences between raw and normalized classification outcomes. Table 3 shows the classification accuracies and AUC values for unnormalized and normalized data using XGBoost[23] and the Aristotle Classifier.[22] The statistically significant difference between the unnormalized and normalized AUC values calculated using DeLong's test is also shown.

For classification by XGBoost, the best-performing strategies had the greatest improvements in classification accuracy and noticeably better AUC values. These include removeBatchEffect, EigenMS, and Cyclic Loess normalization. EigenMS

https://doi.org/10.1021/jasms.3c00295
J. Am. Soc. Mass Spectrom. 2023, 34, 2775−2784

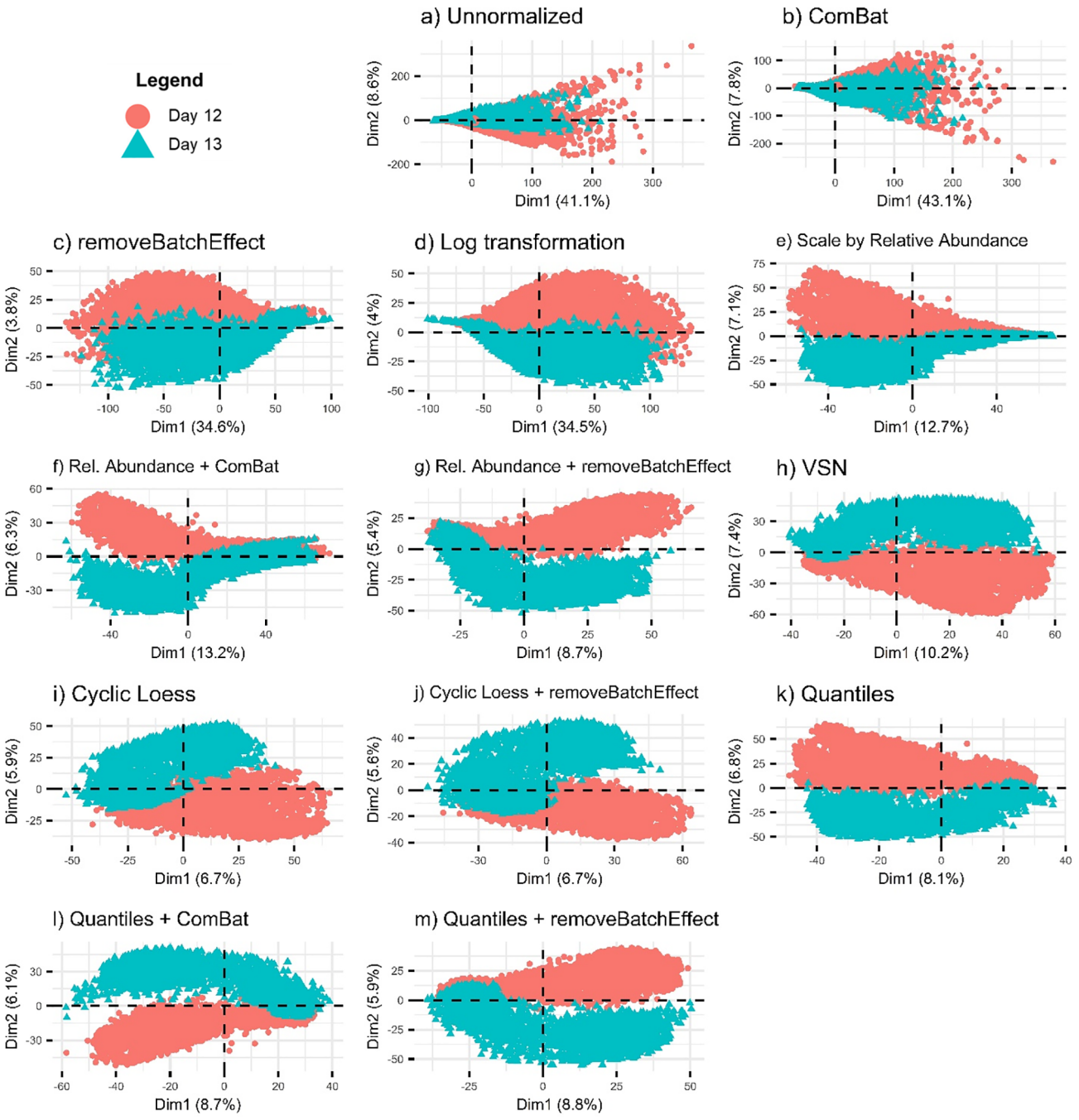


Figure 3. PCA plots for the cancer spheroid dataset (a) before and (b−m) after normalization. Colors indicate different sample types (spheroids harvested on day 12 or day 13).

had the greatest AUC, with a value of 0.9391. The worst-performing method was VSN, with a significant deterioration in AUC and the worst accuracy. All other methods resulted in insignificant changes in the AUC and similar accuracies relative to the unnormalized data.

Upon classification with the Aristotle Classifier, VSN and WaveICA are identified as the worst-performing methods due to significant deterioration in both AUC values and accuracy. The log transformation performed similarly to that of the unnormalized data. All other methods result in a significant improvement in classification outcomes, with an approximate 15−26% improvement in classification accuracy and an increase of ∼0.16−0.30 in normalized AUC values. The best-performing strategy for this classifier is EigenMS, with the highest classification accuracy and AUC value.

Overall Assessment of Performance. To identify which strategy or set of strategies are optimal for this dataset, the results of supervised and unsupervised classification were combined. Three methods (VSN, log transformation, and WaveICA) are poor strategies for the dataset due to increased PCA clustering by batch and poor classification outcomes. Scale by relative abundance, quantiles normalization, ComBat, and all variants of removeBatchEffect and ComBat are considered moderately good strategies. While they perform well in supervised classification by the Aristotle Classifier, these methods either result in increased PCA batch clustering or

https://doi.org/10.1021/jasms.3c00295
J. Am. Soc. Mass Spectrom. 2023, 34, 2775−2784

Table 4. Supervised Classification Outcomes for MALDI-MSI Dataset[a]

Approach	XGBoost			Aristotle Classifier			
	Accuracy (%)	AUC	DeLong P-Value	Accuracy (%)	AUC	DeLong P-value	Performance
Unnormalized	95.65	0.9969	-	92.02	0.9897	-	-
Log Transformation	96.66	0.9969	0.9366	94.22	0.9936	8.07×10^{-34}	Good
Quantiles + removeBatchEffect	95.66	0.9933	1.04−08	98.98	0.9997	7.61×10^{-53}	Good
Relative Abundance + ComBat	96.16	0.9941	5.52×10^{-07}	97.32	0.9981	8.39×10^{-38}	Good
Relative Abundance + removeBatchEffect	96.16	0.9952	0.0005	97.24	0.9983	3.04×10^{-40}	Good
VSN	96.12	0.9941	5.73×10^{-08}	97.20	0.9990	4.40×10^{-53}	Good
Quantiles + ComBat	95.48	0.9919	1.37×10^{-13}	98.21	0.9995	4.20×10^{-51}	Mediocre
Cyclic Loess + removeBatchEffect	94.01	0.9895	4.47×10^{-22}	95.38	0.9916	0.0230	Mediocre
Relative Abundance	93.20	0.9901	1.37×10^{-17}	97.52	0.9988	1.08×10^{-51}	Mediocre
ComBat	93.87	0.9825	9.47×10^{-40}	95.64	0.9946	1.62×10^{-21}	Mediocre
Quantiles	87.73	0.9769	7.24×10^{-54}	95.80	0.9990	1.34×10^{-50}	Mediocre
Cyclic Loess	85.56	0.9670	9.71×10^{-81}	97.7	0.9986	5.15×10^{-52}	Mediocre
removeBatchEffect	97.82	0.9980	0.0033	84.74	0.9522	6.70×10^{-108}	Mediocre

[a]Approaches with accuracies superior to the unnormalized data for both classifiers are identified as "good" strategies, while approaches that had accuracies inferior to the unnormalized data for at least one classifier are indicated as "mediocre."

insignificant differences in XGBoost classification, indicating that normalization is suboptimal. Ultimately, due to significant improvements in supervised classification and evident reductions in PCA clustering by batch, EigenMS, removeBatchEffect, and Cyclic Loess are identified as the best-performing methods for normalizing MS data of extracted lipids acquired from latent fingerprint samples.

Normalization of a Cancer Spheroid Dataset of Metabolites Ionized by MALDI-MSI. As a second demonstration of the normalization assessment strategy, the same workflow was applied to a cancer spheroid MALDI-MSI dataset. EigenMS and WaveICA normalization were omitted from this demonstration due to either an extensive execution time or the lack of relevance for MALDI data; only 12 normalization methods were evaluated. In this dataset, the biological variable is spheroid age, with the spheroids being harvested after either Day 12 or Day 13. A total of ten spheroids at each time point were collected, leading to a total of ten different batches for each spheroid age.

Metric 1: Sample Clustering in Unsupervised Classification. The results of the unsupervised classification are shown in Figure 3. PCA of the unnormalized data reveals distinct sample clustering by biological conditions rather than by batch (Figure 3a). Effective normalization methods would thus see an enhanced separation of the clusters by biological condition, or sample type. The poorest-performing method was ComBat normalization, where the normalized plot mirrors the unnormalized data (Figure 3b). For the remaining 11 methods an evident separation of clusters is observed, indicating good performance for this metric (Figure 3c−m). The dramatic difference before and after normalization underscores the importance of effective data normalization procedures. The potential biological implications of these findings are significant. Without effective data normalization, there is a risk of misinterpreting the biological phenomena under study. Such misinterpretations could lead researchers to draw inaccurate conclusions, which may subsequently influence the direction of further research or even clinical decision-making.

Metric 2: Supervised Classification Outcomes. The results of supervised classification by XGBoost and the Aristotle Classifier are listed in Table 4. For classification by the Aristotle Classifier, all approaches except removeBatchEffect, result in significant improvements in normalized AUC values and classification accuracy. Interestingly, removeBatchEffect is the sole strategy with a significant improvement in normalized AUC and classification accuracy for XGBoost classification. All other methods have AUC values lower than the unnormalized data. Some of these methods (log transformation, quantile + removeBatchEffect, relative abundance + ComBat, relative abundance + removeBatchEffect, and VSN), however, have improvements in classification accuracy. As these methods have accuracies superior to the unnormalized data across both classifiers, they are categorized as good-performing methods. The remaining seven methods are identified as "mediocre" methods due to inferior classification performance for at least one classifier.

Overall Assessment of Performance. By combining the findings from these two metrics, we find that ComBat and removeBatchEffect underperform for normalizing the cancer spheroid dataset due to inferior performance in at least one metric. The other ten methods result in distinct clusters based on spheroid age, which demonstrate good performance for the PCA metric. While these methods also perform well for the second metric, one strategy outperforms the rest. The best-performing strategy for normalizing this cancer spheroid dataset is Quantiles + removeBatchEffect due to significant improvements in supervised classification outcomes via the Aristotle Classifier. Overall, these findings demonstrate that better-performing normalization methods lead to improved classification outcomes.

Impact of the Classifier. Classification outcomes depend on the relationship between the data and the classifier. Some datasets may be better suited for classification by certain tools than others. For instance, when classifying the unnormalized ESI-MS dataset of extracted lipids, XGBoost outperformed the Aristotle Classifier, achieving accuracies of 82.18% compared to the Aristotle Classifier's 62.87% (Table 3). Thus, to achieve good classification outcomes, it may be necessary to evaluate the performance of different supervised classification tools on unnormalized data. Classifiers with good outcomes can subsequently be incorporated into this workflow, potentially replacing XGBoost or the Aristotle Classifier.

Additionally, the performance of a particular normalization strategy appears to depend on the classifier employed due to differences in how the classifiers function. This is evident when comparing the effects of cyclic loess normalization on the

https://doi.org/10.1021/jasms.3c00295
J. Am. Soc. Mass Spectrom. 2023, 34, 2775−2784

MALDI-MSI dataset. While cyclic loess normalization led to an increase in classification accuracy using the Aristotle Classifier, a deterioration in accuracy was seen for XGBoost classification (Table 4). These results highlight the importance of evaluating the classifier and the normalization strategy in tandem. As different normalization strategies impact the overall performance of the classifiers differently, evaluating the performance of a strategy across multiple classifiers can provide deeper insights.

CONCLUSION

We provide a simple and straightforward workflow to facilitate the identification of the best-performing normalization strategies for MS datasets, using supervised and unsupervised machine learning tools. We demonstrate the utility of the workflow by identifying a set of well-performing approaches for normalizing a latent fingerprint dataset of lipids acquired by ESI-MS and a cancer spheroid dataset of metabolites ionized by MALDI-MSI. Using only two key evaluation metrics, this workflow provides researchers the ability to try any existing or emerging normalization strategies on any type of omics data, enabling a simple and unbiased evaluation of normalization performance. With the application of this workflow, optimal approaches can be chosen, ensuring that systematic variability is adequately addressed, thereby increasing the accuracy and reliability of the downstream analysis.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jasms.3c00295.

MSNORM_samplecode (TXT)
Input matrices (ZIP)

AUTHOR INFORMATION

Corresponding Author

Heather Desaire − *Department of Chemistry, University of Kansas, Lawrence, Kansas 66045, United States;* orcid.org/0000-0002-2181-0112; Phone: 785-864-3015; Email: hdesaire@ku.edu

Authors

Aleesa E. Chua − *Department of Chemistry, University of Kansas, Lawrence, Kansas 66045, United States;* orcid.org/0009-0003-7067-3032

Leah D. Pfeifer − *Department of Chemistry, University of Kansas, Lawrence, Kansas 66045, United States*

Emily R. Sekera − *Department of Chemistry and Biochemistry and the Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0002-1668-3227

Amanda B. Hummon − *Department of Chemistry and Biochemistry and the Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0002-1969-9013

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.3c00295

Author Contributions

Experiments: L.P., E.R.S., A.E.C.; Data Analysis: A.E.C.; Manuscript Writing: A.E.C., E.R.S., H.D. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by NIH Grant RF1AG072760 to H.D. and by funding from NSF CHE-1950293 and the University of Kansas Madison and Lila Self Graduate Fellowship to A.E.C.

REFERENCES

(1) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic acids research* **2009**, *37*, W652−W660.

(2) 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* (4), 447−449.

(3) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical chemistry* **2006**, *78* (3), 779−787.

(4) *MetaboAnalyst 5.0 User Statistics*. https://www.metaboanalyst.ca/docs/UserStats.xhtml (accessed June 2, 2023).

(5) Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Orešič, M. Algorithms and tools for the preprocessing of LC-MS metabolomics data. *Chemometrics and Intelligent Laboratory Systems* **2011**, *108* (1), 23−32.

(6) Välikangas, T.; Suomi, T.; Elo, L. L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in bioinformatics* **2018**, *19* (1), 1−11.

(7) Benedetti, E.; Gerstner, N.; Pučić-Baković, M.; Keser, T.; Reiding, K. R.; Ruhaak, L. R.; Štambuk, T.; Selman, M. H.; Rudan, I.; Polašek, O.; et al. Systematic evaluation of normalization methods for glycomics data based on performance of network inference. *Metabolites* **2020**, *10* (7), 271.

(8) Willforss, J.; Chawade, A.; Levander, F. NormalyzerDE: online tool for improved normalization of omics expression data and high-sensitivity differential expression analysis. *J. Proteome Res.* **2019**, *18* (2), 732−740.

(9) Li, B.; Tang, J.; Yang, Q.; Li, S.; Cui, X.; Li, Y.; Chen, Y.; Xue, W.; Li, X.; Zhu, F. NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic acids research* **2017**, *45* (W1), W162−W170.

(10) Yang, Q.; Wang, Y.; Zhang, Y.; Li, F.; Xia, W.; Zhou, Y.; Qiu, Y.; Li, H.; Zhu, F. NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. *Nucleic Acids Res.* **2020**, *48* (W1), W436−W448.

(11) Wen, B.; Mei, Z.; Zeng, C.; Liu, S. metaX: a flexible and comprehensive software for processing metabolomics data. *BMC bioinformatics* **2017**, *18*, 1−14.

(12) Pfeifer, L. D.; Patabandige, M. W.; Desaire, H. Leveraging R (LevR) for fast processing of mass spectrometry data and machine learning: Applications analyzing fingerprints and glycopeptides. *Frontiers in Analytical Science* **2022**, *2*, 961592.

(13) Friedrich, J.; Seidel, C.; Ebner, R.; Kunz-Schughart, L. A. Spheroid-based drug screen: considerations and practical approach. *Nat. Protoc.* **2009**, *4* (3), 309−324.

(14) Ahlf Wheatcraft, D. R.; Liu, X.; Hummon, A. B. Sample Preparation Strategies for Mass Spectrometry Imaging of 3D Cell Culture Models. *JoVE* **2014**, No. 94, No. e52313.

(15) Li, H.; Hummon, A. B. Imaging Mass Spectrometry of Three-Dimensional Cell Culture Systems. *Anal. Chem.* **2011**, *83* (22), 8794−8801.

(16) Tobias, F.; Hummon, A. B. Lipidomic comparison of 2D and 3D colon cancer cell culture models. *Journal of Mass Spectrometry* **2022**, *57* (8), No. e4880.

(17) Hua, D.; Liu, X.; Go, E. P.; Wang, Y.; Hummon, A. B.; Desaire, H. How to Apply Supervised Machine Learning Tools to MS Imaging

Files: Case Study with Cancer Spheroids Undergoing Treatment with the Monoclonal Antibody Cetuximab. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (7), 1350−1357.
(18) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43* (7), No. e47.
(19) Leek, J. T.; Johnson, W. E.; Parker, H. S.; Jaffe, A. E.; Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **2012**, *28* (6), 882−883.
(20) Karpievitch, Y. V.; Nikolic, S. B.; Wilson, R.; Sharman, J. E.; Edwards, L. M. Metabolomics data normalization with EigenMS. *PLoS One* **2014**, *9* (12), No. e116221.
(21) Deng, K.; Zhang, F.; Tan, Q.; Huang, Y.; Song, W.; Rong, Z.; Zhu, Z. J.; Li, K.; Li, Z. WaveICA: A novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. *Anal. Chim. Acta* **2019**, *1061*, 60−69.
(22) Hua, D.; Desaire, H. Improved Discrimination of Disease States Using Proteomics Data with the Updated Aristotle Classifier. *J. Proteome Res.* **2021**, *20* (5), 2823−2829.
(23) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; San Francisco, California, August 13−17, 2016; pp 785−794. DOI: 10.1145/2939672.2939785.
(24) Zhang, Y.; Parmigiani, G.; Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform* **2020**, *2* (3), lqaa078.
(25) Desaire, H. How (Not) to Generate a Highly Predictive Biomarker Panel Using Machine Learning. *J. Proteome Res.* **2022**, *21* (9), 2071−2074.

https://doi.org/10.1021/jasms.3c00295
J. Am. Soc. Mass Spectrom. 2023, 34, 2775−2784