

Review

Instrumental Drift in Untargeted Metabolomics: Optimizing Data Quality with Intrastudy QC Samples

Andre Märten^{1,2} , Johannes Holle³ , Brit Mollenhauer^{4,5}, Andre Wegner¹ , Jennifer Kirwan⁶ 
and Karsten Hiller^{1,*}

¹ Department of Bioinformatics and Biochemistry, Braunschweig Integrated Centre of Systems Biology, Technische Universität Braunschweig, 38118 Braunschweig, Germany

² Physikalisch-Technische Bundesanstalt, 38116 Braunschweig, Germany

³ Department of Pediatric Gastroenterology, Nephrology and Metabolic Diseases, Universitätsmedizin Berlin, 13353 Berlin, Germany

⁴ Department of Neurology, University Medical Center Göttingen, 37073 Göttingen, Germany

⁵ Paracelsus-Elena-Klinik, 34128 Kassel, Germany

⁶ Berlin Institute of Health at Charité, Universitätsmedizin Berlin, 10117 Berlin, Germany

* Correspondence: karsten.hiller@tu-braunschweig.de; Tel.: +49-531-391-55201

Abstract: Untargeted metabolomics is an important tool in studying health and disease and is employed in fields such as biomarker discovery and drug development, as well as precision medicine. Although significant technical advances were made in the field of mass-spectrometry driven metabolomics, instrumental drifts, such as fluctuations in retention time and signal intensity, remain a challenge, particularly in large untargeted metabolomics studies. Therefore, it is crucial to consider these variations during data processing to ensure high-quality data. Here, we will provide recommendations for an optimal data processing workflow using intrastudy quality control (QC) samples that identifies errors resulting from instrumental drifts, such as shifts in retention time and metabolite intensities. Furthermore, we provide an in-depth comparison of the performance of three popular batch-effect correction methods of different complexity. By using different evaluation metrics based on QC samples and a machine learning approach based on biological samples, the performance of the batch-effect correction methods were evaluated. Here, the method TIGER demonstrated the overall best performance by reducing the relative standard deviation of the QCs and dispersion-ratio the most, as well as demonstrating the highest area under the receiver operating characteristic with three different probabilistic classifiers (Logistic regression, Random Forest, and Support Vector Machine). In summary, our recommendations will help to generate high-quality data that are suitable for further downstream processing, leading to more accurate and meaningful insights into the underlying biological processes.



Citation: Märten, A.; Holle, J.; Mollenhauer, B.; Wegner, A.; Kirwan, J.; Hiller, K. Instrumental Drift in Untargeted Metabolomics: Optimizing Data Quality with Intrastudy QC Samples. *Metabolites* **2023**, *13*, 665. <https://doi.org/10.3390/metabo13050665>

Academic Editors: Nicole Strittmatter and Regina Verena Taudte

Received: 1 March 2023

Revised: 8 May 2023

Accepted: 12 May 2023

Published: 16 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: metabolomics; quality control; analytical variation; batch effects

1. Introduction

The metabolome is a collection of small molecules (<1500 Da) or metabolites that are involved in cellular processes, including energy production, signaling, and synthesis of cellular components. It integrates information from the genome and regulatory processes, as well as environmental factors such as diet and lifestyle. Because the metabolome reflects the downstream effects of these factors on cellular function, it is very close to the actual phenotype of a cell, tissue, or organism. As such, studying the metabolome can provide valuable insights into disease mechanisms, biomarker discovery, drug development, and precision medicine [1,2].

The most employed analytical techniques in metabolomics are gas chromatography (GC) and liquid chromatography (LC) coupled to mass spectrometry (MS), both enabling the simultaneous assessment of many metabolites in large cohorts [3,4]. Despite significant

technical advances in the field of mass spectrometry, technical variation still remains a challenge, in particular in large clinical cohort studies. Multiple factors introduce technical variation, which is categorizable in pre-analytical and analytical variation (Figure 1).

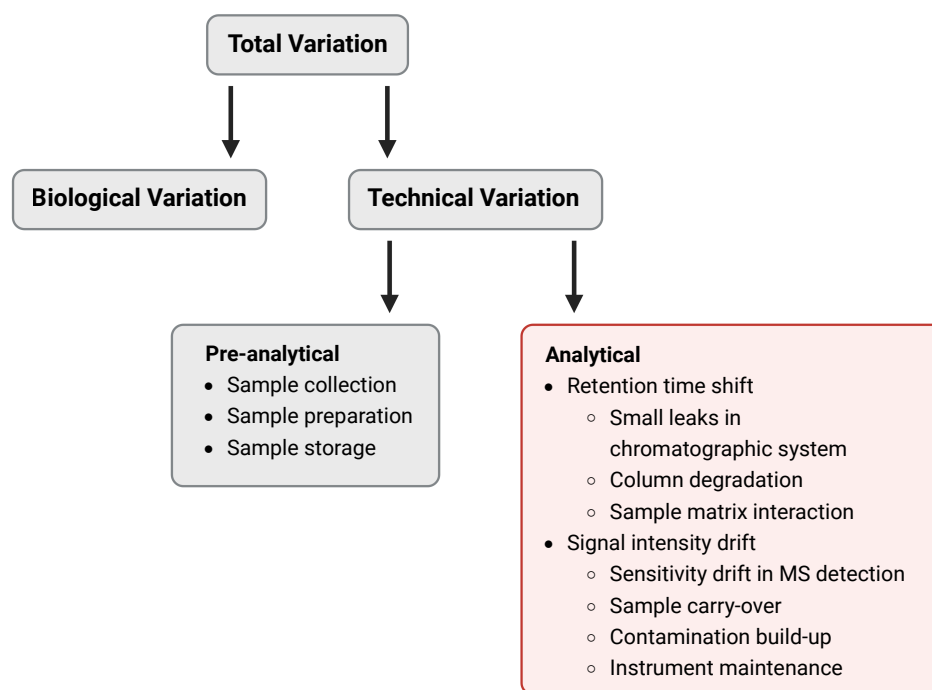


Figure 1. Sources of variation in quantitative metabolomics. The total variation of each data value is comprised of biological, random, and technical variation. Technical variation can be divided in pre-analytical and analytical variation. Pre-analytical variation is induced by either poor methodology or variation in processing during sample collection, preparation or storage. Analytical variation in sample values originates from the analytical technique itself and is reflected in retention time and signal intensity shifts.

Pre-analytical variation is introduced by different collection containers, pre-storage preparation, and sample storage conditions. For example, metabolic profiles of blood plasma can be affected by different types of anticoagulant-coated tubes [5,6] as well as different temperatures and time periods of storage due to conversions and degradation of metabolites [7–9]. However, even when pre-analytical processing is optimal, technical variation introduced by the analytical platform due to instrumental drifts, can never be excluded (analytical variations). The instrumental drifts during both GC- and LC-MS measurements lead to shifts in retention time (RT) [10–16] and signal intensity [17,18], both of which need to be considered during data preprocessing. RT is widely used for aligning chromatographic peaks in GC- and LC-MS runs that belong to identical analytes in different samples [19–21]. External calibration can be incorporated in the metabolome analysis workflow for analyzing a large number of samples, where RT calibrant runs are carried out every 30 to 40 sample runs to calibrate sample chromatograms between the RT calibrant runs [22,23]. This procedure works most of the time, but there are occasional occurrences, such as small leaks in the chromatography system, minor degradation of column performance, and interactions between different compounds of the analyzed sample matrix, where RT shifts may happen in some samples between the RT calibrants.

The occurrence rate of these unexpected RT shift events increases as the number of analyzed samples increases.

For this task, several alignment algorithms and computer programs are publicly and commercially available, but due to the high complexity of the metabolome, further improvements of existing approaches as well as manual interventions are still needed [22,24,25].

We will introduce one strategy to detect and avoid misalignments due to RT shifts in large studies as part of this review.

Another challenge for large cohorts are significant variations of feature intensities due to instrument drifts; these effects are usually categorized into intra- and inter-batch effects. In this regard, a batch is defined as a set of samples processed and analyzed by the same experimental procedure (same operator and instrument) in an uninterrupted manner. Since the capacity of certain chromatographic equipment (e.g., columns, liners) is limited, cohorts with a higher number of samples are typically partitioned into several batches [26]. Intra- and inter-batch effects occur due to multiple reasons, and a significant source of signal variability is a sensitivity drift over time and across batches in MS detection, as metabolite quantification relies on the intensities of MS peaks. In GC-MS, instrumental causes of changes in intensity between batches mainly occur due to instrument maintenance, ageing, and tuning [27,28]. Of course, machine maintenance, such as the exchange of the liner or column cleaning, is necessary to maintain adequate peak intensities in large cohorts. Another source of technical variation is the problem of sample carry-over and contamination build-up, which could differ between batches [17,29,30]. Sample carry-over is caused by samples containing large amounts of metabolites, from which residuals remain on the column and may affect metabolite signals in later samples of the sequence run. Contamination build-up is caused by compounds trapped in the ion source leading to reduced mass spectral performance. These systemic variations cause detectable differences between samples, which can lead to false discoveries, as batch effects can be stronger than inter-phenotype effects, as highlighted in various studies [18,31–33]. Several methods have been developed to tackle this problem, as batch effects in metabolomic experiments are impossible to entirely eliminate. The simplest approach is the randomization of samples within the sequence run. Complete randomization removes the risk of introducing bias, and the variance observed in each biological sample will be a combination of the biological and technical variation. However, when the sample size is so large that the measurement has to be divided into different batches, a blocked approach has to be performed, where only samples within one batch can be compared, reducing the statistical power [34]. Therefore, batch-effects need to be eliminated so that samples between batches can be directly compared. One approach is spiking the samples with labeled internal standards (LIS) as controls. However, in untargeted metabolomics, where all metabolites are of interest, a large number of LIS needs to be added to the samples. This would increase the risk of LIS coeluting with metabolites of interest. Moreover, the added standards may not be representative for the specific chemical characteristics of the unknowns, and response factors may differ. Therefore, spiking with LIS is usually avoided in untargeted metabolomics [26]. The most used and robust methods include the modeling of the above-described batch effects based on intrastudy quality control (QC) samples [21,26,35].

In this review, we outline the importance of incorporating QC samples into the measurement sequence. The review then compares three different methods for adjusting batch effects using QC samples. The first method is a simple and easy-to-implement median-based normalization technique [36], the second method incorporates a regression-based normalization method using a penalized cubic smoothing spline called Quality Control-Robust Spline Correction (QC-RSC) [37]. Finally, a recently published normalization method, Technical variation elimination with ensemble learning architecture (TIGER) [38], is discussed. Our article provides a comparative evaluation of these three strategies using two GC-MS and one LC-MS-based data sets previously recorded in our labs [36,39]. At last, we introduce an effective method for identifying and correcting quantification errors due to peak misalignment. Overall, this review provides a comprehensive guide for researchers to process and analyze untargeted metabolomics data acquired for high sample number cohorts.

2. Intrastudy QC-Samples in Metabolomics

The application of intrastudy QC samples has been recognized as a valuable tool to significantly improve the validity of large-scale metabolomics studies [21,26,35,40,41]. The QC samples should reflect the aggregated metabolite composition of all biological samples for a certain study [40–42]. Typically, the best way to prepare QC samples is to mix all biological test samples in equal amounts [42], because as such, the QCs are closest to the biological samples in means of composition (same sample matrix and metabolites). If the amount of material is limited or sample preparation starts before the last sample has been collected, it is not possible to generate sufficient QCs. To mitigate this, commercially available QC samples can mimic the composition of the biological samples, although to a lower accuracy compared to intrastudy QC samples [43–45]. Dunn et al. used commercially available serum samples, but had to remove 20% of all features, due to differences in metabolic composition between the commercial QC sample and samples from the study population. Therefore, preparing intrastudy QC samples from a representative subset of biological samples is the better solution [44]. Another option is generating intrastudy QCs from the same sample type but from another biological source. A clear disadvantage in this case is that the metabolite concentrations and the sample matrix differ from the biological samples. As a last option, artificial QCs can be created with chemical standards, whereas as many metabolites from as many metabolite classes as possible are dissolved in a dummy matrix [46].

There are three major reasons to employ QC samples: The first is the initial equilibration of the measurement system. Each sequence of samples should start with conditioning QC samples. In the case of GC-MS and LC-MS, usually the data of the first four to eight injections are not stable [47–49]. This effect occurs especially in the context of preventative maintenance, after which active sites of the column are not equilibrated or blocked with the sample matrix. Multiple injections of the QCs prior to the main sample acquisition will condition the column sufficiently [44,47,50]. The actual number of required conditioning samples depends on several factors, particularly sample type, chromatographical system, injection volume, chromatographic column, and mass spectrometric design. It has been suggested that each laboratory should determine an individual optimal number of conditioning samples by injecting up to 50 intrastudy QC samples until reproducible results will be acquired [46]. It is to note that the only intention of the conditioning samples is for column conditioning and not for later batch-effect correction.

The second reason to employ QC samples is the evaluation of measurement precision. As all QCs are equal in terms of metabolite concentration and sample matrix, quantitative quality criteria such as the relative standard deviation (RSD) and Dispersion-ratio (D-ratio) can be determined for quality assessment.

The third and most important reason is the modeling and correction of systemic error. Since QC samples are measured intermittently throughout the whole sequence run, changes in instrument performance can be accurately monitored. Data of QCs quantitatively reveal gradual changes in instrument sensitivity, which is extremely useful for the elimination of batch variations. For this, it is crucial to include a sufficient number of QCs into the sequence to maximize the performance of batch-effect correction and lower the risk of overfitting [51]. On the other hand, injecting too many QCs will significantly extend analysis time, which results in an even more pronounced instrumental drift [26,52]. Kamleh et al. quantitatively evaluated the effect of QC frequency on the reproducibility of metabolic features. The number of reproducible features (RSD < 30% and <15%) was only 1.5% and 5% lower when comparing data corrected with a QC injection every 10th sample with data corrected with a QC injection every fifth sample [53]. Therefore, one QC sample should be injected for every third to 10th biological sample [44,53]. Additionally, it is recommended to append two QCs each at the beginning and end of the sequence to avoid extrapolation during batch-effect correction in case of injection failure [40]. If both QCs run successfully, only one of these QCs is used for batch-effect correction, e.g., the first and the last one.

3. Methods to Correct Metabolomics Data for Batch Effects

Besides just detecting gradual changes in instrument sensitivity, QC samples are also used for the correction of these effects. In general, such a correction is performed as follows: As batch effects are metabolite specific, each recorded metabolite level needs to be analyzed and corrected, separately. The QC recordings for each metabolite define a pattern of instrument-related signal changes as a function of the injection order. Because all biological samples are flanked by corresponding QCs, it can be assumed that instrument-related signal alterations in the QCs also apply to the neighboring biological samples. This is the reason why a mathematical model can be employed to predict batch variation based on the information of the QCs. The predicted batch variability is then subtracted from the original data to yield batch-effect-free data. A qualified QC-based correction does not only account for inter- and intra-batch effects but is also resistant to overfitting. This is important in order to deter the model from accounting for random variations in the data, leading to overoptimistic quality measures, but non-usable data. In the following, we review and compare three often applied correction methods in regard to their effectiveness for batch-effect removal.

3.1. Median Normalization

The first batch-effect correction method is the simplest to apply and just normalizes each sample metabolite signal ($x_{i,sample}$) by the corresponding median signals of neighbouring QCs ($x_{i,QC}$). Specifically, the three in terms of acquisition time and chronologically closest QC samples for each biological sample are chosen. The median signal for each metabolite i is calculated based on these three QC samples ($\bar{x}_{i,QC}$) and is then applied to normalize the metabolite signal of the sample ($x_{i,sample}$). Using the median instead of the mean makes this approach more robust against outliers or missing (zero) values within the QCs (Figure 2A).

3.2. Quality Control-Robust Spline Correction

Quality Control-Robust Spline Correction (QC-RSC) is an advanced regression-based method [37]. For this method, an unweighted cubic spline f is fitted to the QC data (x_{QC}) as a function of the injection order (t_{QC}), with n being the length of t_{QC} (Equation (1)). In contrast to non-parametric models, QC-RSC has the advantage of accounting for more complex batch-related variations in metabolite signals. Furthermore, compared to the Quality Control-Robust LOESS Signal Correction algorithm [44], QC-RSC is computationally more efficient, by replacing the two-step LOESS QC fitting and piece-wise polynomial regression stage with a single-step adaptive cubic smoothing spline algorithm.

The spline f minimizes the distance between model fit and QCs under consideration of a roughness penalty, controlled by the smoothing parameter p ($0 < p < 1$). The roughness penalty penalizes the variability in the function f , with $p \rightarrow 0$ resulting in an interpolating spline and $p \rightarrow 1$ in a linear least squares regression.

$$p \sum_{i=1}^n (x_{QC}(i) - f(t_{QC}(i)))^2 + (1 - p) \int \left(\frac{d^2 f}{dt^2_{QC}} \right)^2 dx \quad (1)$$

To avoid overfitting, p is optimized using leave-one-out cross-validation. Each metabolite is then normalized by its own correction function, which removes intra-batch effects. To remove inter-batch effects, the value is furthermore divided by the median signal of the metabolite of the intrastudy QC samples (Figure 2B).

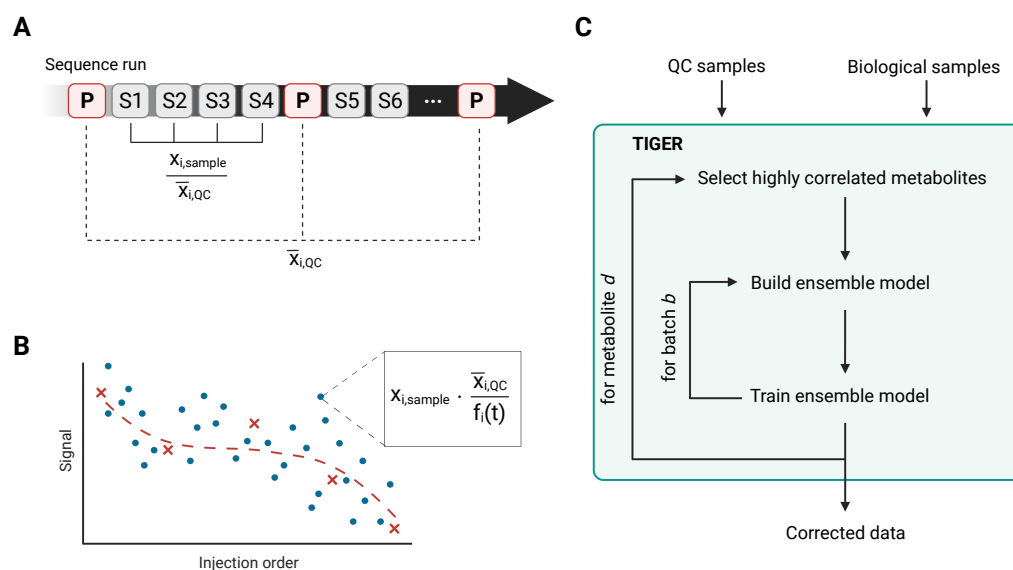


Figure 2. (A) Graphical representation of the median-based normalization. For each sample, the median of the three nearest pools is calculated for each metabolite i . Then, the metabolite intensity for metabolite i is divided by the corresponding median $\bar{x}_{i,QC}$ of the pools. (B) Schematic functionality of the QC-RSC algorithm. For a given metabolite peak, batch effects can be visualized by plotting the metabolite signal against the injection order. Here, the blue circles represent the biological samples and the red crosses represent the QC samples with which the unweighted cubic smoothing spline is fitted (red dashed line). Then, each sample is normalized by multiplication with the correction factor. The correction factor is the quotient of the median signal of the QCs and the value given by the cubic smoothing spline at injection order t of the sample to be corrected. (C) Schematic representation of the TIGER algorithm. The TIGER algorithm can be described in three steps. 1. Variable selection, 2. model construction, and 3. data correction.

3.3. Technical Variation Elimination with Ensemble Learning Architecture

TIGER (Technical variation elimination with ensemble learning architecture) is the most sophisticated algorithm for the batch-effect correction discussed here. It is an adaptable ensemble learning architecture comprised of several base models [38]. The TIGER algorithm starts by selecting metabolites highly correlated with the objective metabolite, which will be the features for the ensemble model. The ensemble model is constructed for each batch separately and consists by default of n Random Forest (RF) models with different hyperparameter combinations, chosen from a hyperparameter pool defined by the user. The Random Forest model is trained on data comprised of error ratios y' of the objective metabolite and the raw metabolite signals of the correlated metabolites X . The error ratio is calculated as follows:

$$y' = \frac{y - \bar{y}}{\bar{y}} \quad (2)$$

Here, y denotes the raw signal of the objective metabolite and \bar{y} the median of y across the whole data set. The model's performance is evaluated in a K -fold cross validation with a loss function of $L(y, y')$,

$$L(\hat{y}, y') = \frac{1}{K} \sum_{k=1}^K \frac{|\hat{y}^{(k)} - y'^{(k)}|}{y'^{(k)}} \quad (3)$$

where $\hat{y}^{(k)}$ is the predicted error ratio and $y'^{(k)}$ the actual error ratio of the k th CV fold. Based on the loss function, the base model will receive a weight, such that high-performing models

have high weights and under-performing models have low weights, but its information is still considered.

$$w_i = \frac{\exp(-L(\hat{y}_i, y'))}{\sum_n^i \exp(L(\hat{y}_i, y'))} \quad (4)$$

Here, n is the number of all base models, and i is the i th base model. For the actual data correction, the base models are retrained on the whole data set. The error ratios y' and metabolite signals X are used to train the RF model. Hence, the RF model will also predict error ratios, which need to be converted back to metabolite signals. The final result of the algorithm is the weighted sums of all base models (Figure 2C).

By selecting only a small set of features and the RF algorithm, TIGER's base models are of moderate complexity, thus mitigating the risk of capturing random noise in addition to the technical variation. Moreover, an ensemble learning architecture is employed, further improving the models' robustness and lowering the risk of overfitting by considering the output from strong as well as weak models. Altogether, the TIGER algorithm prioritizes robustness and high generalization over high complexity, making this a valuable method even for data with small sample sizes or data with a high degree of noise.

4. Evaluation of Batch-Effect Correction Methods

4.1. Evaluation Metrics

To evaluate the performance of the different batch-effect correction methods, we employed three different quantitative quality criteria, namely the RSD and D-ratio for each metabolite, as well as the Euclidean distance of QCs after the principal component analysis (PCA). The RSD is a widely used metric, which is calculated for each metabolite i within the QCs by dividing the standard deviation $\sigma_{i,QC}$ by the arithmetic mean $\bar{m}_{i,QC}$. This leads to a unitless and standardized measure comparable among all detectable metabolites.

$$RSD_{i,QC} = \frac{\sigma_{i,QC}}{\bar{m}_{i,QC}} \cdot 100\% \quad (5)$$

A typically accepted RSD threshold for metabolites in biomarker discovery should be below 20% for LC-MS and below 30% for GC-MS [18,33,44,47,48]. However, only observing the RSD may result in over-optimistic results, as batch-effect correction methods remove the batch effects based on the QCs. This way, an over correction of the data, which could lead to the removal of the biological variation, would not be detected. Thus, we additionally applied the relation of the statistical dispersion of the QCs to the dispersion of the biological test samples to evaluate the normalization performance [54,55]. The D-ratio is calculated by dividing the technical variation by the total observed variation; this is the sum of technical and biological variation. Here, the variance of the QCs ($\sigma_{i,QC}^2$) approximates the technical variation, and the variance of the biological test samples ($\sigma_{i,sample}^2$) approximates the overall biological variation [46].

$$D-ratio_i \approx \frac{\sqrt{\sigma_{i,QC}^2}}{\sqrt{\sigma_{i,sample}^2 + \sigma_{i,QC}^2}} \cdot 100\% \quad (6)$$

A D-ratio close to 0% would be a perfect measurement, where the technical variance is zero and all observed variance originates from the biological variation. On the other hand, a D-ratio of 100% would be the worst possible measurement, where there is no biological variation and only noise is detected. A metabolite, where $\sigma_{i,sample}^2 \gg \sigma_{i,QC}^2$ with a D-ratio below 50%, is preferred [55].

Another method to visually evaluate batch-effect-removal is the PCA. By plotting the first two principal components, the clustering of batches and the removal of those can be observed. High quality data show tightly clustered QC data points at the origin of the PCA and equally distributed sample points around the QCs. The Euclidean distance between

the centroid of the QCs, and each QC sample point can be calculated to quantitatively assess the effect of the batch-effect correction on the QCs itself.

$$d(C_{QC}, x_{i, QC}) = \sqrt{(C_{QC, PC1} - x_{QC, PC1})^2 + (C_{QC, PC2} - x_{QC, PC2})^2} \quad (7)$$

Here, $d(C_{QC}, x_{i, QC})$, represents the Euclidean distance between the centroid of all QC samples C_{QC} and each individual QC sample $x_{i, QC}$ in the PCA. Each sample point as well as the centroid have two coordinates, which are denoted as the principal components $PC1$ and $PC2$, respectively. The Euclidean distance of QCs is frequently used to evaluate the analytical variability [30]. These three performance measures provide an in-depth picture for evaluating the batch-effect correction methods.

4.2. Comparison of Batch-Effect Correction Methods

To provide recommendations on which method to choose, we evaluated all of the three above highlighted batch-effect correction methods and performed RSD, D-ratio, and a PCA. Furthermore, the methods were applied to normalize GC-MS and LC-MS data sets to evaluate differences in measurement techniques.

All batch-effect correction methods reduced the technical error in each tested data set, as demonstrated by the reduced RSD as compared to the raw data (Figure 3A). Here, TIGER performed best for each data set with a reduction of the median RSD by more than 60%. QC-RSC reduced the median RSD by more than 50% and the median normalization by more than 40% for each data set. Furthermore, all methods were able to reduce the D-ratio (Figure 3B). Again, TIGER performed best by reducing the median D-ratio by 71%, 64%, and 43% for GC-MS 1 and 2 data and the LC-MS data, respectively. For the GC-MS 1 data, QC-RSC performed better than the median normalization, with a reduction of the D-ratio by 41% compared to 27%. For GC-MS data 2 and the LC-MS data, QC-RSC and the median normalization performed equally good by reducing the D-ratio by approximately 50%.

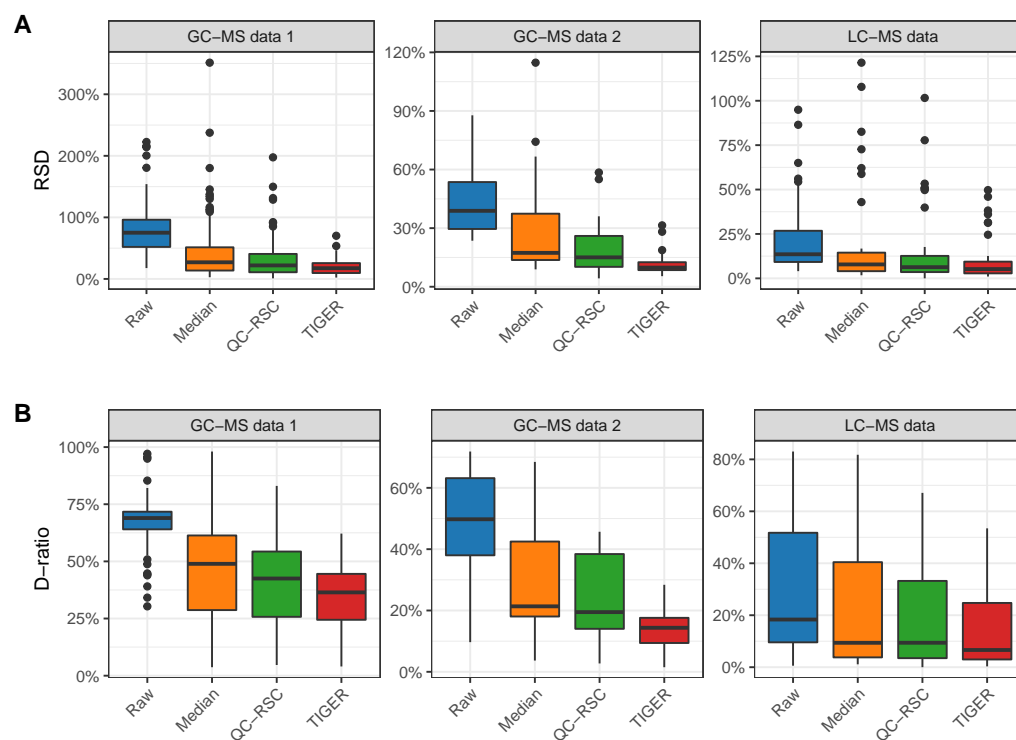


Figure 3. Comparison of normalization methods. A total of 140 metabolic features are present in GC-MS data 1, 25 in GC-MS data 2, and 42 in the LC-MS data. (A) Distribution of RSD of QC samples before and after normalization. (B) Distribution of D-ratio before and after normalization. Outliers are shown as black dots.

To further evaluate the different methods, we performed a PCA. As described in the previous chapter, we expected all QC samples to cluster tightly together after removing the batch effects, while we expected all biological samples to be distributed across the plot. Figure 4 depicts the PCA plots for the raw data and the batch-effect correction methods.

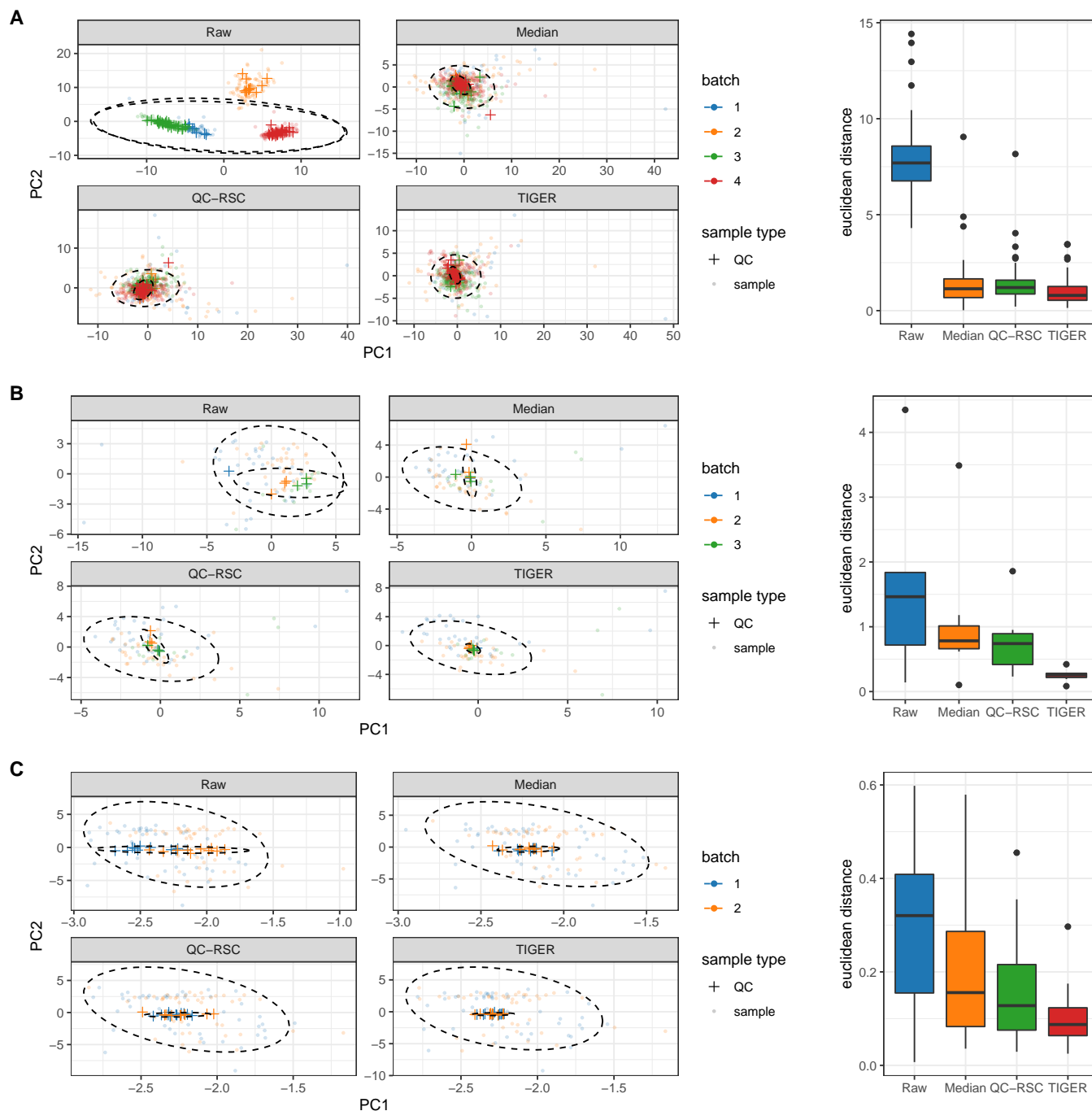


Figure 4. PCA plots of different normalization methods for (A) GC-MS data 1, (B) GC-MS data 2, and (C) LC-MS data. QCs are represented as crosses and biological samples as dots, which are partially transparent. Box plots show the euclidean distance of the QCs to its centroid within the PCA plots. Outliers are shown as black dots.

The PCA of the raw data clearly partitioned the QCs into the underlying batches for every data set, with GC-MS 1 having the most substantial batch effects. After batch-effect correction, all methods resulted in similar plots with tightly clustered QC samples at the origin of the PCA plot scores and equally distributed biological samples. Therefore, all

methods were able to eliminate batch effects regardless of the analytical technique and led to a small analytical variation relative to the biological variation. To quantitatively analyze the PCA plots, we calculated the Euclidean distance between each QC sample point and its corresponding centroid (Figure 4). The Euclidean distance was reduced with all batch-effect correction methods for each data set compared to the raw data. Here, QC-RSC and the median normalization perform similarly for all data sets with a median euclidean distance of approximately 1.2 for GC-MS 1, 0.75 for GC-MS 2, and 0.13 for LC-MS data. The TIGER normalization achieves the lowest distance, which is slightly better than QC-RSC and median normalization. All methods were able to reduce the technical error without overfitting and keeping the biological variation intact by reducing both RSD and D-ratio of the QC samples. Additionally, the removal of batch effects could be observed in the PCA plots and by calculating the Euclidean distances of the QCs. For all quality criteria, TIGER outperformed all other methods.

However, it is important to note, that the previously described evaluation criteria are all based on the same QCs used to fit and train the batch-effect correction methods. This could lead to overoptimistic results; furthermore, it does not measure the impact of the batch-effect correction on the biological samples. For this purpose, different machine learning classifiers were trained to classify persons into Parkinson's disease or healthy control groups based on a published cerebrospinal fluid metabolomics data set to test the power of these algorithms in terms of biological information [36]. Before training the machine learning classifiers, the three batch-effect correction methods had been applied to normalize the data based on intermittently recorded QCs. The stratification performance of these classifiers indicates whether a batch-effect correction method reduces or increases the predictability of phenotypes by interfering biological information. After correction for batch effects by the three methods, we at first removed metabolites that were not accepted by our quality criteria (RSD < 30% and D-ratio < 50%). For the median normalization, 52 metabolites passed the quality acceptance criteria (37.1% of all detected metabolites), 66 for QC-RSC (47.1% of all detected metabolites), and 103 for TIGER (73.6% of all detected metabolites).

Next, we selected the metabolic features to train the model on using recursive feature elimination (RFE). This method trains the model on the full number of features and assigns each feature an importance metric. The least important features are removed, and the process is repeated until a pre-defined number of features will be reached. Here, ten features were selected for each corrected data set by RFE, resulting in three sets of features. Citramalate, pyroglutamate, tryptophan, urea, and glycine were present in all three feature sets and, finally, were used to train the classification models. Here, we focus on three popular probabilistic classifiers, namely boosted Logistic Regression (LogitBoost), Random Forest (RF), and radial kernel Support Vector Machine (svmRadial). For the optimal evaluation of the classifiers, we performed a double repeated cross-validation (CV) approach with two loops. The outer loop generates 100 random splits of training and hold-out sets. The inner loop is used to tune the models' hyperparameters on the training data by maximizing the Area under the Receiver Operating Characteristic curve (AUROC) with repeated CV (10 repeats, 5 folds). The model with the highest AUROC in the CV then predicts the patients' class of the hold-out set. The results of the predictions is shown in Figure 5. Based on the ROC curve evaluation, the models trained on the TIGER corrected data reached the highest AUROC for each classifier (0.979 ± 0.021 for LogitBoost, 0.969 ± 0.009 for RF, 0.958 ± 0.016 for svmRadial), followed by the models trained on the data corrected with QC-RSC (0.963 ± 0.019 for LogitBoost, 0.893 ± 0.025 for RF, and 0.889 ± 0.026 for svmRadial). The worst performance was observed for the classifiers trained on the median-corrected data (0.760 ± 0.008 for LogitBoost, 0.841 ± 0.011 for RF, and 0.820 ± 0.021 for svmRadial).

In summary, TIGER performed best for the batch-effect correction in our evaluation reflected in the lowest RSD, D-ratio, and median Euclidean distance of QC samples in the PCA. This was also the reason why many more metabolites passed our predefined quality criteria and remained in the data set for further analysis. In addition, the performance of

machine learning classifiers trained on the data set corrected with TIGER demonstrated the best performance, emphasizing that this batch-effect correction method captures the inter-phenotype information present in the data in the most optimal way. All in all, we recommend using TIGER as the preferred method for the batch effect correction due to the minor loss of metabolic information. On the other hand, QC-RSC is in the advantage for the normalization of very large cohorts, as the computation costs of TIGER are relatively high.

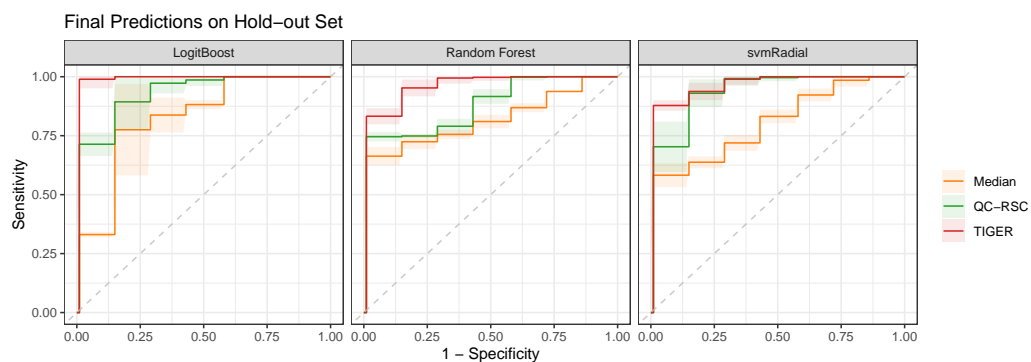


Figure 5. Results of machine learning classifier on corrected data sets. The performance of three probabilistic classifiers were evaluated on predicting the phenotype based on the metabolic signature of CSF samples. The classes of Parkinson’s disease and the healthy control had to be predicted. For all three classifiers, the models trained on the TIGER-corrected data achieved the highest AUROC (0.979 ± 0.021 for LogitBoost, 0.969 ± 0.009 for RF, and 0.958 ± 0.016 for svmRadial), followed by the models trained on data corrected by QC-RSC (0.963 ± 0.019 for LogitBoost, 0.893 ± 0.025 for RF, and 0.889 ± 0.026 for svmRadial). The models trained with the median-corrected data demonstrated the worst performance (0.760 ± 0.008 for LogitBoost, 0.841 ± 0.011 for RF, and 0.820 ± 0.021 for svmRadial).

5. Advanced Strategies to Further Improve Metabolite Quantification and Chromatogram Alignment

A prerequisite for the successful application of the above discussed batch-effect correction algorithms is an accurate chromatogram alignment and picking of quantification peaks. The correct alignment of metabolite features across measurements is crucial for an accurate quantification in the context of every metabolomics’ analysis. RT shifts of molecular features within- and between batches can result in wrong alignments of metabolite features, especially for closely eluting metabolites with similar or even identical mass spectra, such as isomers. In the following, we present a strategy to screen and to correct for quantification errors due to metabolite feature misalignments. We take advantage of the fact that EI ionization generates highly reproducible fragmentation patterns for a certain metabolite. We assume that an increased or decreased metabolite amount in the sample affects all fragment ions equally, and that the same applies for a potential drop in instrument sensitivity over time. Hence, the ratio of a pair of fragment ion intensities must be identical for a certain metabolite over all measurements and independent of the sample or instrument condition. As a demo data set, we chose a previously published GC-MS-based metabolomics data set that was recorded for CSF samples in the context of Parkinson’s disease [36].

Within this data set, gluconic acid elutes at 25.65 min and the peak integrals of fragment ions 205, 305, and 333 were automatically assigned for quantification (Figure 6A). However, two coeluting compounds produce ion chromatographic signals on SIC 205 at 25.5 and 25.88 as well (Figure 6B). Due to the close elution of these metabolites, there is a substantial risk of picking the wrong peaks of this SIC for integration during automatic data processing. Integrating the wrong peak of this SIC results in either lower or higher log-fold changes of ion ratios, which can be indicative for a misalignment if laying outside the determined outlier threshold (Figure 6C). In such a case, outlier values can either be removed from the data set or imputed based on the quantification of the other metabolite ions (here

305 and 333). Depending on the type of missing values, different imputation methods should be chosen for optimal results. Here, the values are missing at random (MAR), due to suboptimal data preprocessing. Therefore, we chose the Random Forest imputation method that performed best for MAR [56,57]. For a detailed evaluation of imputation methods with further types of missing values, we refer the reader to the paper of Wei et al. [57]. By taking advantage of this approach, we identified 48 falsely picked peaks for gluconic acid out of 600 total measurements (Figure 6D). The imputation of metabolite signals based on the other metabolite ions decreased the RSD for this metabolite by 12%.

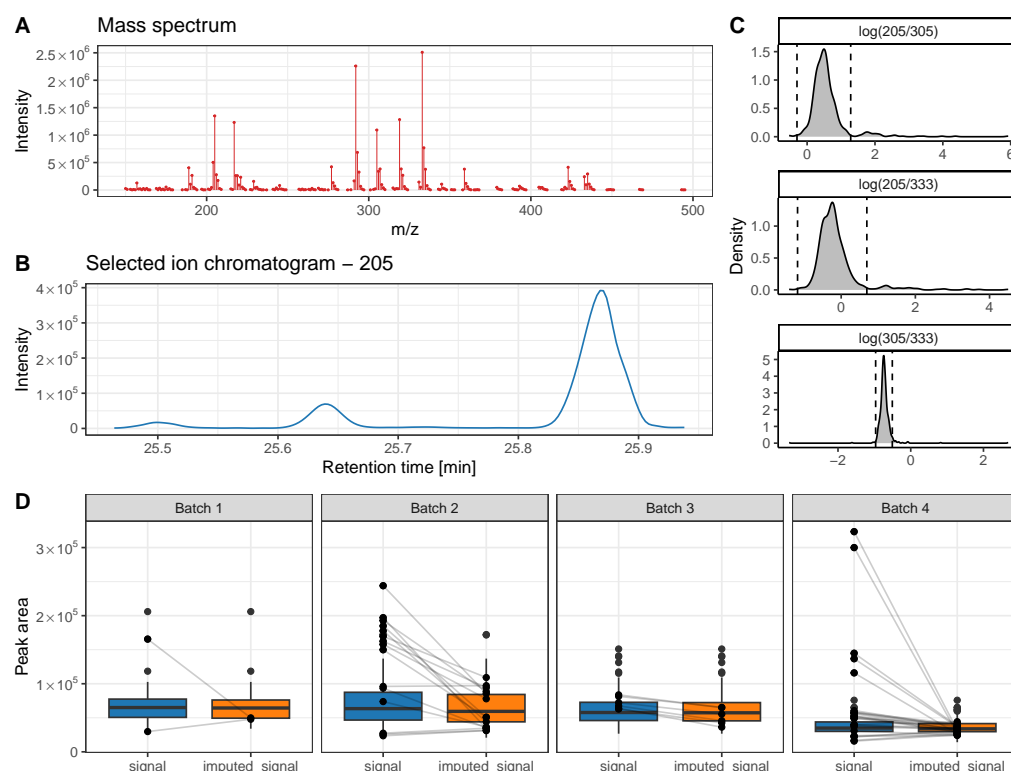


Figure 6. Identification of quantification errors in chromatograms of gluconic acid. (A) Mass spectrum. (B) Selected ion chromatogram of mass 205. (C) Distribution of quantification ion ratios. Vertical dotted lines depict the threshold of accepted values, which is calculated by the median signal intensity \pm three times the MAD. (D) Box plots of raw metabolite signals and imputed signals, based on outlier detection.

6. Conclusions and Future Directions

In this article, we reviewed and evaluated three popular algorithms suited for the elimination of mass spectrometric noise based on intermittently measured QC samples in metabolomics studies with a high number of samples. Furthermore, we introduced a strategy to improve chromatogram alignment and peak picking in GC-MS data. To increase the quality of large metabolomics studies, an optimal workflow should include the following steps: 1. Preparation of adequate QCs is essential for bigger cohorts. 2. The sequence run should start with five to ten QC samples solely for equilibration of the analytical system. 3. The sequence of sample measurements should be random and at least one QC sample should be measured in-between three to seven samples. In addition there should be two QC samples at the sequence beginning and end. 4. For high sample numbers, the sample blocks between QCs should be increased to avoid an unnecessary extension in analysis time. For lower sample numbers, the sequence should contain at least eight QCs (without the conditioning QCs). After data acquisition, the peak picking, integration and mapping of metabolites should be verified. For this, we propose an easy-to-implement method to check the quantification of metabolites based on log-fold changes

of the quantification ions' intensities. To advise which method to employ for the batch effect removal, we evaluated three different methods: a median-based approach, QC-RSC, and TIGER. Although all three methods significantly removed batch effects and drifts in instrument sensitivity, TIGER always outperformed the other two methods. For this reason and despite the high computational cost of this algorithm, we advise using TIGER. Many more metabolites passed the quality criteria ($RSD < 30\%$ and $D\text{-ratio} < 50\%$), and overall separation between tested phenotypes was more evident, as highlighted in the better performance of employed machine learning classifiers.

Untargeted metabolomics is an important tool for biomarker discovery, drug development, and precision medicine. These fields rely heavily on large data sets to provide the needed statistical power. Here, the machine learning approach with an ensemble learning architecture has been proven to be the most promising tool for batch-effect correction. Therefore, the performance of batch-effect correction could be further improved by applying deep learning approaches. Furthermore, this could enable the comparison of untargeted metabolomic studies between instruments, which opens the possibility of mining large databases across studies. Potentially, this could lead to the identification of biological mechanisms or biomarkers that would otherwise be hidden in smaller individual studies.

Author Contributions: Conceptualization, A.M. and K.H.; software, A.M.; formal analysis, A.M.; resources, K.H.; methodology, A.M., J.H., B.M., K.H. and J.K.; data curation, A.M.; writing—original draft preparation, A.M. and K.H.; writing—review and editing, A.M., J.H., B.M., A.W., J.K. and K.H.; visualization, A.M.; supervision, K.H.; project administration, A.M. and K.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors thank Dominik Müller for giving access to the GC-MS 2 data set used in this review. Figure 2 was created with BioRender.com, (accessed on 8 May 2023).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUROC	Area under the Receiver Operating Characteristic
CSF	Cerebrospinal fluid
CV	Cross validation
D-ratio	Dispersion-ratio
GC	Gas chromatography
LC	Liquid chromatography
LogitBoost	Boosted Logistic Regression
MS	Mass spectrometry
PCA	Principal component analysis
QC	Quality control
QC-RSC	Quality Control-Robust Spline Correction
RF	Random Forest
RFE	Recursive feature elimination
ROC	Receiver Operating Characteristic
RSD	Relative standard deviation
RT	Retention time
svmRadial	Radial Kernel Support Vector Machine
TIGER	Technical variation elimination with ensemble learning architecture

References

1. Xia, J.; Broadhurst, D.I.; Wilson, M.; Wishart, D.S. Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics* **2013**, *9*, 280–299. [[CrossRef](#)] [[PubMed](#)]
2. Wishart, D.S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* **2016**, *15*, 473–484. [[CrossRef](#)] [[PubMed](#)]
3. Mapstone, M.; Cheema, A.K.; Fiandaca, M.S.; Zhong, X.; Mhyre, T.R.; MacArthur, L.H.; Hall, W.J.; Fisher, S.G.; Peterson, D.R.; Haley, J.M.; et al. Plasma phospholipids identify antecedent memory impairment in older adults. *Nat. Med.* **2014**, *20*, 415–418. [[CrossRef](#)] [[PubMed](#)]
4. Siskos, A.P.; Jain, P.; Römisch-Margl, W.; Bennett, M.; Achaintre, D.; Asad, Y.; Marney, L.; Richardson, L.; Koulman, A.; Griffin, J.L.; et al. Interlaboratory Reproducibility of a Targeted Metabolomics Platform for Analysis of Human Serum and Plasma. *Anal. Chem.* **2017**, *89*, 656–665. [[CrossRef](#)]
5. Barri, T.; Dragsted, L.O. UPLC-ESI-QTOF/MS and multivariate data analysis for blood plasma and serum metabolomics: Effect of experimental artefacts and anticoagulant. *Anal. Chim. Acta* **2013**, *768*, 118–128. [[CrossRef](#)] [[PubMed](#)]
6. Chen, D.; Han, W.; Su, X.; Li, L.; Li, L. Overcoming Sample Matrix Effect in Quantitative Blood Metabolomics Using Chemical Isotope Labeling Liquid Chromatography Mass Spectrometry. *Anal. Chem.* **2017**, *89*, 9424–9431. [[CrossRef](#)] [[PubMed](#)]
7. Zhou, H.; Yuen, P.; Pisitkun, T.; Gonzales, P.; Yasuda, H.; Dear, J.; Gross, P.; Knepper, M.; Star, R. Collection, storage, preservation, and normalization of human urinary exosomes for biomarker discovery. *Kidney Int.* **2006**, *69*, 1471–1476. [[CrossRef](#)]
8. Ammerlaan, W.; Trezzi, J.P.; Lescuyer, P.; Mathay, C.; Hiller, K.; Betsou, F. Method Validation for Preparing Serum and Plasma Samples from Human Blood for Downstream Proteomic, Metabolomic, and Circulating Nucleic Acid-Based Applications. *Biopreservation Biobanking* **2014**, *12*, 269–280. [[CrossRef](#)]
9. Trezzi, J.P.; Bulla, A.; Bellora, C.; Rose, M.; Lescuyer, P.; Kiehnopf, M.; Hiller, K.; Betsou, F. LacaScore: A novel plasma sample quality control tool based on ascorbic acid and lactic acid levels. *Metabolomics* **2016**, *12*, 96. [[CrossRef](#)]
10. Malmquist, G.; Danielsson, R. Alignment of chromatographic profiles for principal component analysis: A prerequisite for fingerprinting methods. *J. Chromatogr. A* **1994**, *687*, 71–78. [[CrossRef](#)]
11. Duran, A.L.; Yang, J.; Wang, L.; Sumner, L.W. Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* **2003**, *19*, 2283–2293. [[CrossRef](#)]
12. Johnson, K.J.; Wright, B.W.; Jarman, K.H.; Synovec, R.E. High-speed peak matching algorithm for retention time alignment of gas chromatographic data for chemometric analysis. *J. Chromatogr. A* **2003**, *996*, 141–155. [[CrossRef](#)] [[PubMed](#)]
13. Katajamaa, M.; Orešič, M. Processing methods for differential analysis of LC/MS profile data. *BMC Bioinform.* **2005**, *6*, 179. [[CrossRef](#)] [[PubMed](#)]
14. Smith, C.A.; Want, E.J.; O’Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78*, 779–787. [[CrossRef](#)] [[PubMed](#)]
15. De Souza, D.P.; Saunders, E.C.; McConville, M.J.; Likić, V.A. Progressive peak clustering in GC-MS Metabolomic experiments applied to *Leishmania* parasites. *Bioinformatics* **2006**, *22*, 1391–1396. [[CrossRef](#)]
16. Styczynski, M.P.; Moxley, J.F.; Tong, L.V.; Walther, J.L.; Jensen, K.L.; Stephanopoulos, G.N. Systematic Identification of Conserved Metabolites in GC/MS Data for Metabolomics and Biomarker Discovery. *Anal. Chem.* **2006**, *79*, 966–973. [[CrossRef](#)]
17. Burton, L.; Ivosev, G.; Tate, S.; Impey, G.; Wingate, J.; Bonner, R. Instrumental and experimental effects in LC-MS-based metabolomics. *J. Chromatogr. B* **2008**, *871*, 227–235. [[CrossRef](#)]
18. Wang, S.Y.; Kuo, C.H.; Tseng, Y.J. Batch Normalizer: A Fast Total Abundance Regression Calibration Method to Simultaneously Adjust Batch and Injection Order Effects in Liquid Chromatography/Time-of-Flight Mass Spectrometry-Based Metabolomics Data and Comparison with Current Calibration Methods. *Anal. Chem.* **2012**, *85*, 1037–1046. [[CrossRef](#)]
19. Tautenhahn, R.; Patti, G.J.; Rinehart, D.; Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal. Chem.* **2012**, *84*, 5035–5039. [[CrossRef](#)]
20. Zhang, W.; Lei, Z.; Huhman, D.; Sumner, L.W.; and Zhao, P.X. MET-XAlign: A Metabolite Cross-Alignment Tool for LC/MS-Based Comparative Metabolomics. *Anal. Chem.* **2015**, *87*, 9114–9119. [[CrossRef](#)]
21. Brunius, C.; Shi, L.; Landberg, R. Large-scale untargeted LC-MS metabolomics data correction using between-batch feature alignment and cluster-based within-batch signal intensity drift correction. *Metabolomics* **2016**, *12*, 173. [[CrossRef](#)] [[PubMed](#)]
22. Hiller, K.; Hangebrauk, J.; Jäger, C.; Spura, J.; Schreiber, K.; Schomburg, D. MetaboliteDetector: Comprehensive Analysis Tool for Targeted and Nontargeted GC/MS Based Metabolome Analysis. *Anal. Chem.* **2009**, *81*, 3429–3439. [[CrossRef](#)] [[PubMed](#)]
23. Huan, T.; Wu, Y.; Tang, C.; Lin, G.; Li, L. DnsID in MyCompoundID for Rapid Identification of Dansylated Amine- and Phenol-Containing Metabolites in LC-MS-Based Metabolomics. *Anal. Chem.* **2015**, *87*, 9838–9845. [[CrossRef](#)] [[PubMed](#)]
24. Benton, H.P.; Wong, D.M.; Trauger, S.A.; Siuzdak, G. XCMS2: Processing Tandem Mass Spectrometry Data for Metabolite Identification and Structural Characterization. *Anal. Chem.* **2008**, *80*, 6382–6389. [[CrossRef](#)]
25. Clasquin, M.F.; Melamud, E.; Rabinowitz, J.D. *LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2012. [[CrossRef](#)]
26. Wehrens, R.; Hageman, J.A.; van Eeuwijk, F.; Kooke, R.; Flood, P.J.; Wijnker, E.; Keurentjes, J.J.B.; Lommen, A.; van Eekelen, H.D.L.M.; Hall, R.D.; et al. Improved batch correction in untargeted MS-based metabolomics. *Metabolomics* **2016**, *12*, 88. [[CrossRef](#)]

27. Cardinal, M.; Viallon, C.; Thonat, C.; Berdagué, J.L. Pyrolysis-mass spectrometry for rapid classification of oysters according to rearing area. *Endocr. Disruptors* **2000**, *28*, 825–829. [[CrossRef](#)]
28. Pérès, C.; Viallon, C.; Berdagué, J.L. Solid-Phase Microextraction-Mass Spectrometry: A New Approach to the Rapid Characterization of Cheeses. *Anal. Chem.* **2001**, *73*, 1030–1036. [[CrossRef](#)]
29. Fernández-Albert, F.; Llorach, R.; Garcia-Aloy, M.; Ziyatdinov, A.; Andres-Lacueva, C.; Perera, A. Intensity drift removal in LC/MS metabolomics by common variance compensation. *Bioinformatics* **2014**, *30*, 2899–2905. [[CrossRef](#)]
30. Shen, X.; Gong, X.; Cai, Y.; Guo, Y.; Tu, J.; Li, H.; Zhang, T.; Wang, J.; Xue, F.; Zhu, Z.J. Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* **2016**, *12*, 89. [[CrossRef](#)]
31. Boccard, J.; Tonoli, D.; Strajhar, P.; Jeanneret, F.; Odermatt, A.; Rudaz, S. Removal of batch effects using stratified subsampling of metabolomic data for in vitro endocrine disruptors screening. *Talanta* **2019**, *195*, 77–86. [[CrossRef](#)]
32. 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. [[CrossRef](#)] [[PubMed](#)]
33. Zhao, Y.; Hao, Z.; Zhao, C.; Zhao, J.; Zhang, J.; Li, Y.; Li, L.; Huang, X.; Lin, X.; Zeng, Z.; et al. A Novel Strategy for Large-Scale Metabolomics Study by Calibrating Gross and Systematic Errors in Gas Chromatography–Mass Spectrometry. *Anal. Chem.* **2016**, *88*, 2234–2242. [[CrossRef](#)] [[PubMed](#)]
34. Forshed, J. Experimental Design in Clinical ‘Omics Biomarker Discovery. *J. Proteome Res.* **2016**, *16*, 3954–3960. [[CrossRef](#)] [[PubMed](#)]
35. 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 Res.* **2017**, *45*, W162–W170. [[CrossRef](#)]
36. Trezzi, J.P.; Galozzi, S.; Jaeger, C.; Barkovits, K.; Brockmann, K.; Maetzler, W.; Berg, D.; Marcus, K.; Betsou, F.; Hiller, K.; et al. Distinct metabolomic signature in cerebrospinal fluid in early parkinson’s disease. *Mov. Disord.* **2017**, *32*, 1401–1408. [[CrossRef](#)]
37. Kirwan, J.A.; Broadhurst, D.I.; Davidson, R.L.; Viant, M.R. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. *Anal. Bioanal. Chem.* **2013**, *405*, 5147–5157. [[CrossRef](#)]
38. Han, S.; Huang, J.; Foppiano, F.; Prehn, C.; Adamski, J.; Suhre, K.; Li, Y.; Matullo, G.; Schliess, F.; Gieger, C.; et al. TIGER: Technical variation elimination for metabolomics data using ensemble learning architecture. *Brief. Bioinform.* **2022**, *23*, bbab535. [[CrossRef](#)]
39. Holle, J.; Bartolomaeus, H.; Löber, U.; Behrens, F.; Bartolomaeus, T.U.; Anandakumar, H.; Wimmer, M.I.; Vu, D.L.; Kuhring, M.; Brüning, U.; et al. Inflammation in Children with CKD Linked to Gut Dysbiosis and Metabolite Imbalance. *J. Am. Soc. Nephrol.* **2022**, *33*, 2259–2275. [[CrossRef](#)]
40. Dunn, W.B.; Wilson, I.D.; Nicholls, A.W.; Broadhurst, D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012**, *4*, 2249–2264. [[CrossRef](#)]
41. Liu, X.; Zheng, P.; Zhao, X.; Zhang, Y.; Hu, C.; Li, J.; Zhao, J.; Zhou, J.; Xie, P.; Xu, G. Discovery and Validation of Plasma Biomarkers for Major Depressive Disorder Classification Based on Liquid Chromatography–Mass Spectrometry. *J. Proteome Res.* **2015**, *14*, 2322–2330. [[CrossRef](#)]
42. Sangster, T.; Major, H.; Plumb, R.; Wilson, A.J.; Wilson, I.D. A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabolomic analysis. *Analyst* **2006**, *131*, 1075. [[CrossRef](#)] [[PubMed](#)]
43. Barton, R.H.; Nicholson, J.K.; Elliott, P.; Holmes, E. High-throughput 1H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: Validation study. *Int. J. Epidemiol.* **2008**, *37*, i31–i40. [[CrossRef](#)] [[PubMed](#)]
44. Dunn, W.B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J.D.; Halsall, A.; et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **2011**, *6*, 1060–1083. [[CrossRef](#)] [[PubMed](#)]
45. Godzien, J.; Alonso-Herranz, V.; Barbas, C.; Armitage, E.G. Controlling the quality of metabolomics data: New strategies to get the best out of the QC sample. *Metabolomics* **2014**, *11*, 518–528. [[CrossRef](#)]
46. Broadhurst, D.; Goodacre, R.; Reinke, S.N.; Kuligowski, J.; Wilson, I.D.; Lewis, M.R.; Dunn, W.B. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **2018**, *14*, 72. [[CrossRef](#)]
47. Zelena, E.; Dunn, W.B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K.M.; Begley, P.; O’Hagan, S.; Knowles, J.D.; Halsall, A.; Wilson, I.D.; et al. Development of a Robust and Repeatable UPLC-MS Method for the Long-Term Metabolomic Study of Human Serum. *Anal. Chem.* **2009**, *81*, 1357–1364. [[CrossRef](#)]
48. Begley, P.; Francis-McIntyre, S.; Dunn, W.B.; Broadhurst, D.I.; Halsall, A.; Tseng, A.; Knowles, J.; Goodacre, R.; and, D.B.K. Development and Performance of a Gas Chromatography–Time-of-Flight Mass Spectrometry Analysis for Large-Scale Nontargeted Metabolomic Studies of Human Serum. *Anal. Chem.* **2009**, *81*, 7038–7046. [[CrossRef](#)]
49. Michopoulos, F.; Lai, L.; Gika, H.; Theodoridis, G.; Wilson, I. UPLC-MS-Based Analysis of Human Plasma for Metabonomics Using Solvent Precipitation or Solid Phase Extraction. *J. Proteome Res.* **2009**, *8*, 2114–2121. [[CrossRef](#)]
50. Gika, H.G.; Theodoridis, G.A.; Wingate, J.E.; Wilson, I.D. Within-Day Reproducibility of an HPLC-MS-Based Method for Metabonomic Analysis: Application to Human Urine. *J. Proteome Res.* **2007**, *6*, 3291–3303. [[CrossRef](#)]

51. Rong, Z.; Tan, Q.; Cao, L.; Zhang, L.; Deng, K.; Huang, Y.; Zhu, Z.J.; Li, Z.; Li, K. NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2020**, *92*, 5082–5090. [[CrossRef](#)]
52. Kuligowski, J.; Sánchez-Illana, Á.; Sanjuán-Herráez, D.; Vento, M.; Quintás, G. Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC). *Anal. Chem.* **2015**, *140*, 7810–7817. [[CrossRef](#)] [[PubMed](#)]
53. Kamleh, M.A.; Ebbels, T.M.D.; Spagou, K.; Masson, P.; Want, E.J. Optimizing the Use of Quality Control Samples for Signal Drift Correction in Large-Scale Urine Metabolic Profiling Studies. *Anal. Chem.* **2012**, *84*, 2670–2677. [[CrossRef](#)] [[PubMed](#)]
54. Lewis, M.R.; Pearce, J.T.M.; Spagou, K.; Green, M.; Dona, A.C.; Yuen, A.H.Y.; David, M.; Berry, D.J.; Chappell, K.; van der Sluis, V.H.; et al. Development and Application of Ultra-Performance Liquid Chromatography-TOF MS for Precision Large Scale Urinary Metabolic Phenotyping. *Anal. Chem.* **2016**, *88*, 9004–9013. [[CrossRef](#)] [[PubMed](#)]
55. Reinke, S.N.; Gallart-Ayala, H.; Gómez, C.; Checa, A.; Fauland, A.; Naz, S.; Kamleh, M.A.; Djukanović, R.; Hinks, T.S.; Wheelock, C.E. Metabolomics analysis identifies different metabolotypes of asthma severity. *Eur. Respir. J.* **2017**, *49*, 1601740. [[CrossRef](#)] [[PubMed](#)]
56. Stekhoven, D.J.; Buhlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [[CrossRef](#)] [[PubMed](#)]
57. Wei, R.; Wang, J.; Su, M.; Jia, E.; Chen, S.; Chen, T.; Ni, Y. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **2018**, *8*, 663. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.