ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

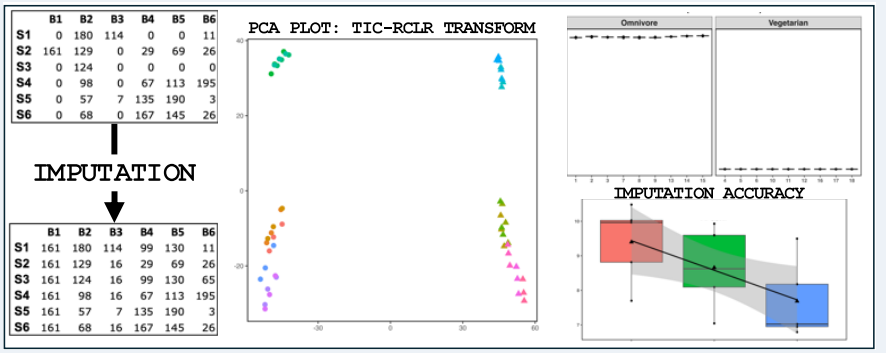
Article

To Impute or Not To Impute in Untargeted Metabolomics—That is the Compositional Question

Dennis D. Krutkin, Sydney Thomas, Simone Zuffa, Prajit Rajkumar, Rob Knight, Pieter C. Dorrestein, and Scott T. Kelley*

Cite This: https://doi.org/10.1021/jasms.4c00434

Read Online



ABSTRACT: Untargeted metabolomics often produce large datasets with missing values. These missing values are derived from biological or technical factors and can undermine statistical analyses and lead to biased biological interpretations. Imputation methods, such as *k*-Nearest Neighbors (kNN) and Random Forest (RF) regression, are commonly used, but their effects vary depending on the type of missing data, e.g., Missing Completely At Random (MCAR) and Missing Not At Random (MNAR). Here, we determined the impacts of degree and type of missing data on the accuracy of kNN and RF imputation using two datasets: a targeted metabolomic dataset with spiked-in standards and an untargeted metabolomic dataset. We also assessed the effect of compositional data approaches (CoDA), such as the centered log-ratio (CLR) transform, on data interpretation since these methods are increasingly being used in metabolomics. Overall, we found that kNN and RF performed more accurately when the proportion of missing data across samples for a metabolic feature was low. However, these imputations could not handle MNAR data and generated wildly inflated or imputed values where none should exist. Furthermore, we show that the proportion of missing values had a strong impact on the accuracy of imputation, which affected the interpretation of the results. Our results suggest imputation should be used with extreme caution even with modest levels of missing data and especially when the type of missingness is unknown.

INTRODUCTION

The exponential growth of untargeted mass spectrometry methods mirrors the growth in sequencing technologies and has generated extraordinary new insights into human and animal physiology,[1,2] disease processes,[3,4] microbial communities,[5] disease biomarkers,[6] and drug discovery.[7] Making sense of these large datasets, including spectral matching, statistical analysis, and machine learning, requires mathematical and statistical approaches to identify data patterns in noisy data to understand the biology. Missing values are quite common in large untargeted metabolomics datasets and can comprise up to 50% of the dataset and affect as many as 80% of the variables.[8,9] There are both biological and technical reasons that values may be absent. A metabolite in a sample may be missing due to (1) an ion of a molecule being absent or below the limit of detection; (2) technical issues such as ion suppression;[10,11] (3) variability in sample processing;[12] (4) variations in stability of molecules;[13,14] or (5) differences in ionization efficiencies.[15]

Missing values compromise the completeness of data which undermines the reliability of both univariate and multivariate statistical analyses such as fold-change analysis, *t*-tests, Analysis of Variance (ANOVA), regression-based analyses, and Principal Component Analysis (PCA).[16] Additionally, missing values result in the loss of critical biochemical information, impeding the identification of patterns, biomarkers, and the understanding of biological pathways and their interac-

Received: October 29, 2024
Revised: February 10, 2025
Accepted: February 13, 2025

tions.[17,18] Because many instances of missing values represent false negatives, dozens of methods for imputing missing values have been developed.[19,20] While variable in their approaches, imputation methods model missing data based on the nonmissing values in the dataset. For instance, methods such as *k*-Nearest Neighbors (kNN) and Random Forest (RF) derive missing metabolite values using the values that exist for that same metabolite in other samples. This suggests that the amount of available data (or inversely the amount of missing data, i.e., "missingness") may have a strong impact on the accuracy of imputation.

Another factor that could significantly impact the accuracy of imputations is the type of missing data which, if incorrectly assumed or modeled, could lead to inaccurate imputation values.[21] In the metabolomics data, researchers have identified three primary classes of missing data: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Each mechanism has distinct characteristics and implications for data analysis and imputation methods.[22,23] Data points are MCAR when the probability of missing is the same for all observations. MCAR data points are a random subset of the complete data, which means that analyses performed on the observed data can be unbiased. Some studies have shown that simple imputation methods, such as minimum, mean, or median imputation, can be appropriate in this case[23−25]; while others found that constant value substitutions offered poor performance, compared to more-sophisticated imputation approaches.[26] In the context of metabolomics, data points would be MCAR in cases where some technical replicates detected signals, while others did not. With simple imputation, when data points are MCAR, median value substitution has previously been shown to be more robust compared to mean value imputation, as it is less sensitive to extreme outliers.[27] Data are MAR when the probability of missingness depends only on the observed data and not on the missing data itself. An example of MAR data within metabolomics would be cases where undetected signals could be explained by the presence or absence of signals for another metabolite (e.g., inability to deconvolute coeluting compounds).[28] When data are MAR, more-sophisticated imputation methods are needed, such as multiple imputation or maximum likelihood methods, that leverage the relationships within the observed data to account for the missingness.[29] Data are MNAR when the probability of missingness is related to external factors outside the features within the dataset. The most common example of MNAR data in metabolomics is the absence of signals at the limit of detection (LOD). With LOD MNAR data, prior studies employed minimum value substitution, or instrument LOD substitution (if known).[30,31] Other MNAR data could be due to differences related to the sample type. MNAR imputation is the most challenging, as it requires assumptions about the unobserved data or external information to model the missingness mechanism.[32] In practice, it is very difficult to distinguish MCAR and MAR,[33] with some researchers advocating for alternative viewpoints for missing data altogether.[34]

An additional important aspect of metabolomics analysis is the issue of compositionality. Metabolomics, especially untargeted metabolomics, can have both compositional and noncompositional character. In many fields, including metabolomics, data may exist in a compositional form, meaning that the data represent parts of a whole. Compositional data are defined as constrained to sum to a fixed value (e.g., 1% or 100%), making the values within one file interdependent. In other words, if one value in a compositional dataset increases, all of the other values must decrease, making the values nonindependent (in technical terms, the data lie in a simplex rather than in Euclidean space). Alleviating this problem requires specialized transformations such as the centered log-ratio (CLR), additive log-ratio (ALR), or isometric log-ratio (ILR), which convert the compositional data to real number space. Traditional statistical methods assume that data points are independent and can vary freely, but the inherent constraints of compositional data create dependencies between the components, leading to potential pitfalls in data interpretation, including spurious correlations and misleading patterns.[35]

In untargeted metabolomics, the absolute value of the peak area can reflect the concentration of an analyte present in the sample. When not normalized to a fixed total, these values exhibit a noncompositional character because they are independent of the presence of other metabolites in the analysis. However, the distinction between compositionality and noncompositionality can sometimes be ambiguous, often depending on how the data are processed and interpreted. For instance, consider a fixed-volume plasma sample, where 100 metabolites are detected. In a different sample, the same 100 metabolites are present, along with a highly abundant drug. If one normalizes each sample to 100%, which is a common practice in untargeted metabolomics, all other metabolite signals in the second sample are proportionally reduced due to the presence of the drug. This normalization does not reflect a true decrease in their absolute quantities, thus introducing a compositional bias, despite the original concentrations remaining unchanged. Another example of this complexity arises from ion suppression, where the introduction of certain molecules can hinder the detection of other metabolites. In this scenario, the measured abundance of some metabolites may be artificially lowered, complicating the interpretation of compositional versus noncompositional data. Further complicating this discussion are the structural versus observational zero values. Structural zeros indicate inherent limits in what is present, influencing the modeling approach and interpretation of results, while observational zeros reflect gaps in instrument sensitivity (peak detection settings) that can skew the perceived composition and relationships among components. Recognizing these impacts helps in the analysis and interpretation of compositional studies.

While raw metabolomic data can comprise a mix of compositional and noncompositional values, metabolite data becomes truly compositional after Total Ion Count (TIC) normalization, also known as relative abundance normalization. TIC normalization is commonly used with metabolomics datasets, particularly in mass-spectrometry based studies.[36,37] TIC normalization corrects for variations in overall signal intensity that may arise from differences in sample loading, ionization efficiency, or instrumental sensitivity during data acquisition.[38] These variations can introduce significant bias and obscure the true biological differences between samples. TIC normalization mitigates these technical artifacts by scaling the signal intensities of individual metabolites relative to the total ion current.[39] Critically, TIC normalization adjusts the raw signal intensities of metabolites relative to the total ion current of the sample, effectively converting the data to a composition. TIC normalization ensures that the metabolite

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

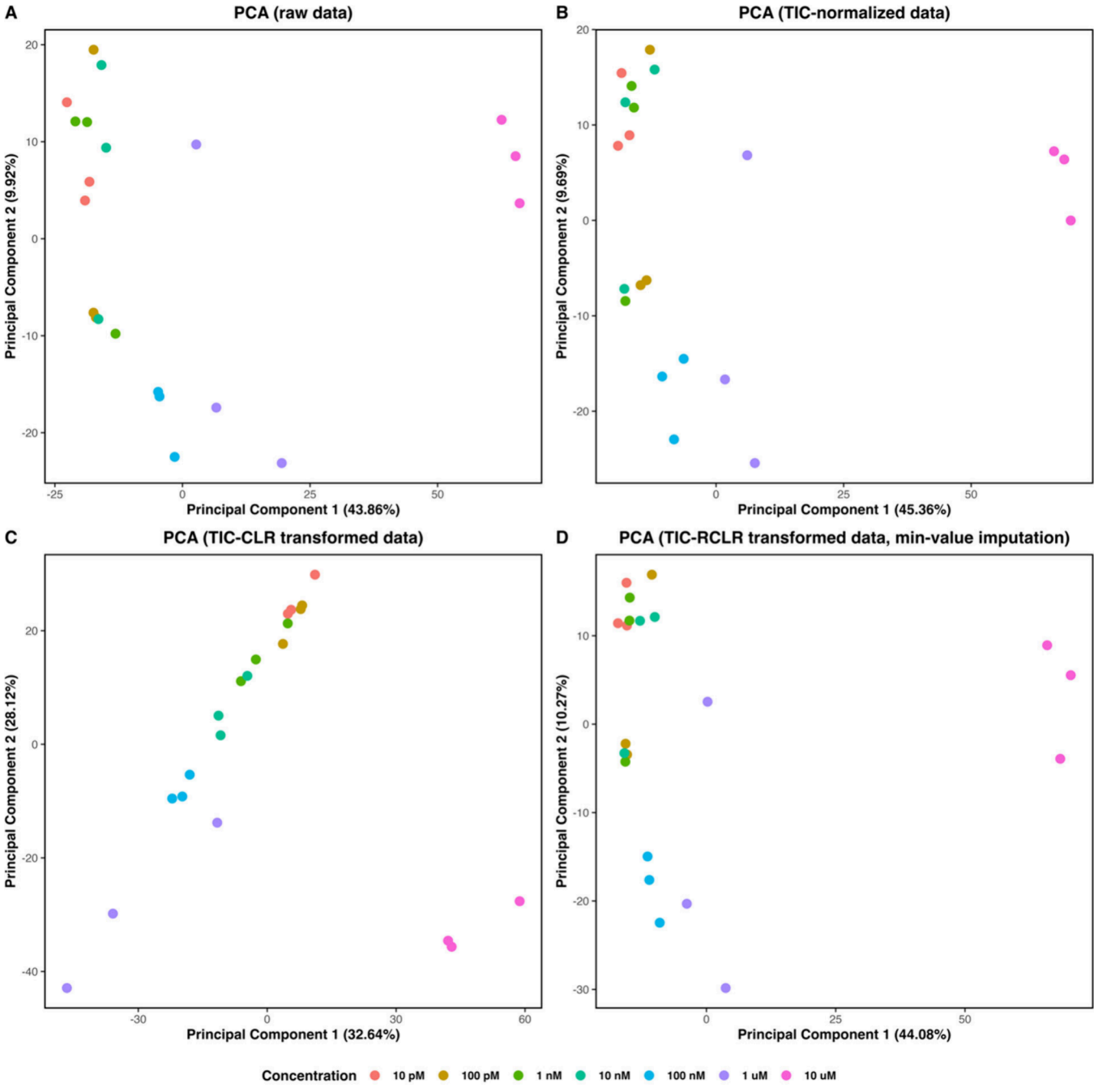


Figure 1. PCA plots for different transformations on targeted metabolomics data: (A) PCA plot for raw data, (B) PCA plot for TIC normalized data, (C) PCA plot for TIC-CLR transformed data (pseudocount = 1×10^{-12}), and (D) PCA plot for TIC-RCLR transformed data, followed by minimum value imputation.

abundances are expressed as relative proportions of the total signal, making them comparable across different samples.

Understanding and applying compositional data analysis (CoDA) in metabolomics are critical because they allow for more-accurate interpretation of the biological significance of relative metabolite abundances. By recognizing the compositional nature of TIC-normalized data, researchers can apply appropriate transformations, such as the CLR, to analyze the data without introducing biases or misinterpretations. The CLR transformation[40] is a straightforward CoDA transformation that can be directly used in numerous statistical analyses and machine learning approaches, and is frequent applied in the fields of geology, molecular biology, and microbial ecology.[41−43] CLR transformation addresses the nonindependence issue of compositional data by converting the data into real number space.[43] With the CLR transformation, each data point in a sample is replaced by the log-ratio of that sample value to the geometric mean, a form of averaging of the data from that sample, ensuring that all data points are centered around a mutual reference point. Another useful application of the CLR transformation for metabolomics arises in its ability to facilitate comparative data analysis. Researchers are increasingly combining data from multiple 'omics approaches (metagenomics, metabolomics, transcriptomics) to determine potential associations among them. Often, however, the distributions and scales of the values

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

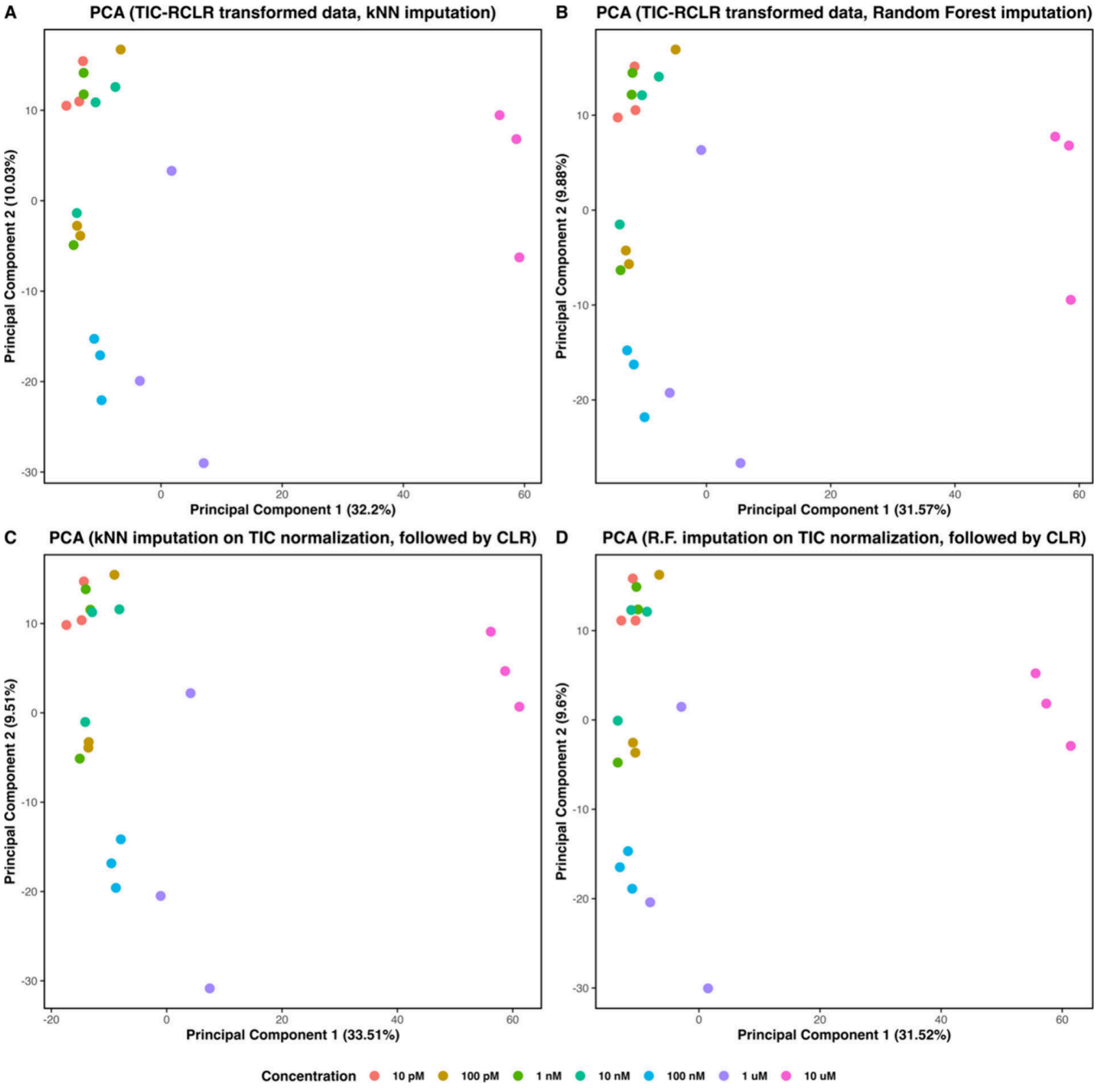


Figure 2. PCA plots of the targeted metabolomics data using machine learning imputations and two different log-ratio transformation methods. (A) TIC-RCLR transformed data, followed by kNN imputation; (B) TIC-RCLR transformed data, followed by RF imputation; (C) kNN imputation on TIC normalized data, followed by CLR transformation; and (D) RF imputation on TIC normalized data, followed by CLR transformation.

among these datasets can be remarkably different, and many statistical tests assume or require similar distributions across data. Using a common transformation ensures that all the datasets to be compared or integrated have a similar distribution and scale for comparisons.

In this study, we examined the effects of missing values on imputation using two of the most commonly used methods, namely, *k*-Nearest Neighbors (kNN) and Random Forest (RF) using two unsimulated metabolomic datasets. The kNN and RF methods are frequently used in the literature and are often included in software packages. They have also been the most extensively validated in prior studies using simulated data.[20] While there are dozens of methods available for imputing data,[44] our goal was not to compare and contrast all the different imputation approaches. Rather, the focus of this article was on the general impact of missing values and compositional data analysis on imputation. Specifically, our goals were to determine (1) how overall levels of missing data generally impact imputation and analytic interpretation of metabolomics data; and (2) how these imputation methods perform with very different types of missing data, namely MNAR and MCAR. Analyses with and without imputation were performed using raw data, TIC normalized data, and two types of CoDA transformation, CLR and the Robust CLR

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

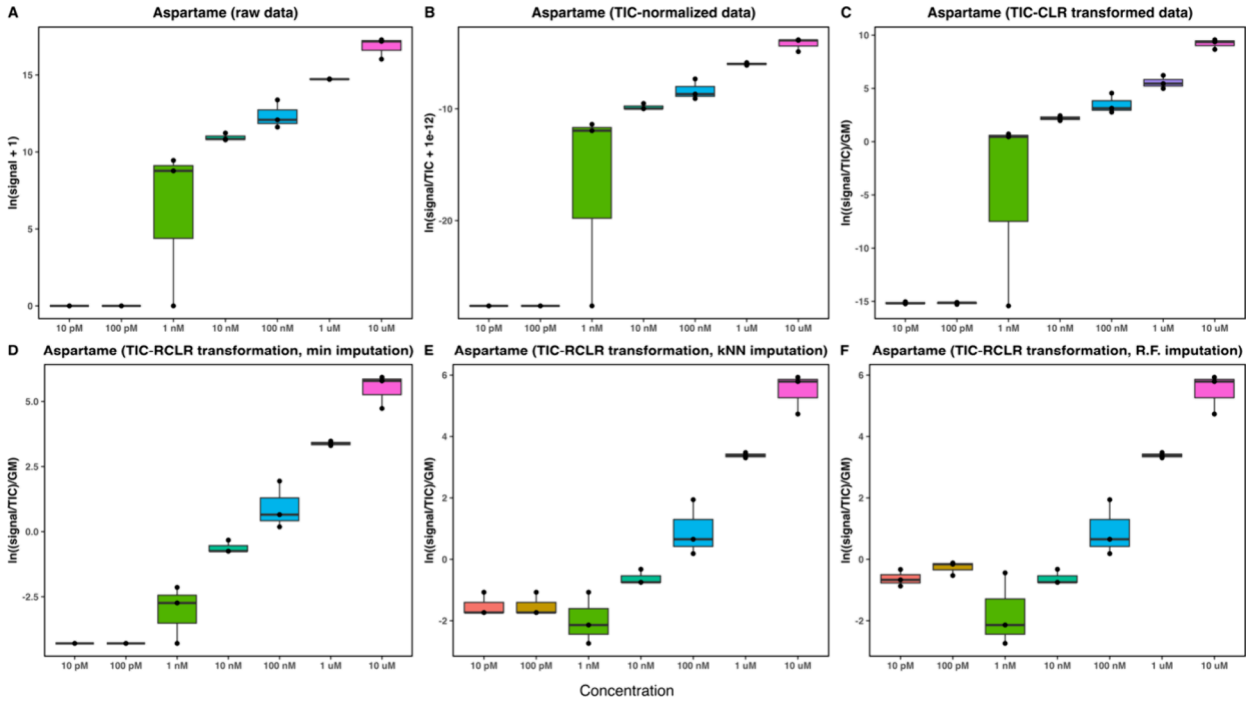


Figure 3. Boxplots of transformations and imputations for aspartame, where the two lowest concentration thresholds have all missing data. (A) Raw data plotted on a natural log scale. (B) TIC normalized data plotted on a natural logarithmic scale. (C) TIC-CLR transformed data (pseudocount = 1×10^{-12}). (D) TIC-RCLR transformed data, followed by minimum value imputation. (E) TIC-RCLR transformed data, followed by kNN imputation. (F) TIC-RCLR transformed data, followed by RF imputation.

(RCLR) transformation,[45] and we examined the impacts of imputation at the whole sample-level and at the level of individual metabolites. Our results showed that both the amount and type of missing data had a profound impact on imputation and downstream interpretation, particularly at the level of individual metabolites, and caution against the blind use of imputation methods without a complete understanding of the types of missing data inherent in a dataset.

RESULTS

Several studies have already evaluated the accuracy of various imputation methods on untargeted metabolomics analysis. These studies largely rely on simulated datasets and with low numbers of missing values. However, most untargeted metabolomics data have a lot of missing values, and thus the results should be evaluated with few to large numbers of missing values. Here, we explore the effectiveness of commonly used imputation methods on two unsimulated datasets. In particular, we assessed how variation between samples changes, examined the effects of imputations when data are MCAR vs MNAR missing at the metabolite level, and evaluated the accuracy of imputations when missing values are introduced randomly and the true value is known. The first dataset was targeted, which included metabolites that were spiked into the samples at fixed concentrations along a logarithmic gradient. The second NIST dataset was untargeted and examined the metabolite profiles of homogenized human fecal samples which had either a vegetarian or omnivorous diet. The effect of the different imputations was assessed in conjunction with common normalization/transformation methods, such as TIC normalization, CLR transformation, and RCLR transformation.

How Do Compositional Transformations and Imputations Affect Targeted Metabolomics Data Where Data is MNAR?

The first dataset we examined had data which was MNAR, because it included metabolites spiked into an untargeted dataset at known concentrations (i.e., not random) that should be detected following the concentration gradient. However, some signals were not detected at the lowest concentrations for a subset of the metabolites due to the limits of detection. Spiked samples are commonly used as internal standards and should not have missing signals, giving motivation to impute these missing values, because we can guarantee the presence of the metabolite at a given concentration. The use of compositional transformations on the TIC normalized data is appropriate because all of the individual ion counts are divided by the TIC, which sums to a fixed value (1.0 or 100%). While compositional analysis of the TIC-normalized data is clearly warranted, we also recognize that researchers directly transform and impute the raw data, often with similar results. Therefore, we have repeated the CoDA below with the raw data and included these results as supplemental data for comparison. For the CoDA analysis, we applied CLR and the newer RCLR transformations. Note, with CLR zeros are replaced prior to transformation with a very small number, while with RCLR zeros are ignored during the transformation but can be replaced with small numbers for downstream analyses. (Zero handling is necessary because the CLR and RCLR transforms both apply the logarithm function, which is undefined at 0.)

To assess the impacts of normalization and compositional transformation on MNAR data, we compared the PCA clustering of samples at increasing concentrations without imputation. Figures 1A–D show PCA analyses for the targeted dataset with no transformations (raw data), TIC normalization, TIC-CLR transformation (TIC normalization followed by CLR), and TIC-RCLR transformation (TIC normalization followed by RCLR) without machine learning imputations (0

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

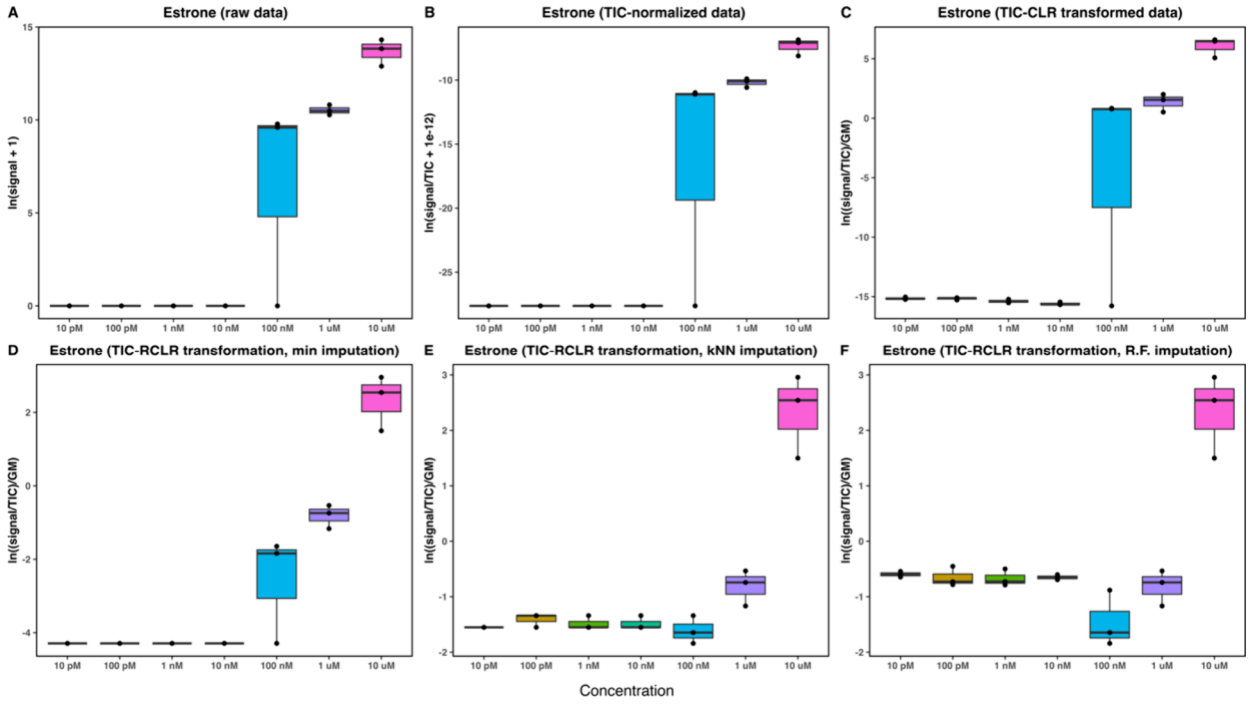


Figure 4. Boxplots of transformations and imputations for estrone, which is missing signals for the four lowest concentrations. (A) Raw data plotted on a natural log scale. (B) TIC normalized data plotted on a natural log scale. (C) TIC-CLR transformed data (pseudocount = 1×10^{-12}). (D) TIC-RCLR transformed data, followed by minimum value imputation. (E) TIC-RCLR transformed data, followed by kNN imputation. (F) TIC-RCLR transformed data, followed by RF imputation.

values substituted with minimum value following transformation). The total variation explained by the different methods was similar. PCA for TIC-CLR transformed data explained the most variation with the first two principal components (60.76%; Figure 1C), while the PCA for the raw data captured the least (53.78%; Figure 1A). PCA for TIC-CLR transformed data captured less variation on the first principal component and more variation on the second principal component compared with other methods (Figure 1C). Clustering of replicates within concentration conditions was similar across the different methods, with the exception of the TIC-CLR transformed data, which contained different clustering patterns compared to those of the other three approaches (Figure 1C). It is worth noting that selecting varying values of pseudocounts for the TIC-CLR transformation produces different clustering of samples. However, since the TIC normalized data contains many proportions with very small values, a pseudocount smaller than all TIC proportions was appropriate (1×10^{-12}). The same analysis was performed after transforming the raw data instead of the TIC-normalized data, which produced similar results, although the CLR explained less of the variation in the first two principal components compared to the TIC-CLR transformation (Figure S1).

To determine the effects of imputation methods on PCA clustering and visualization, we then applied two commonly used imputation methods: kNN and RF. We compared cases where TIC-normalized data was first imputed then CLR transformed versus RCLR-transforming the TIC-normalized data then imputing missing values for the unaffected 0s. Since imputation can be performed either before or after transformation, we examined both possibilities. Figures 2A–D shows PCA analyses for the targeted dataset using machine learning imputations for missing data with TIC-RCLR and TIC-CLR transformations. Since CLR transformation does not accept missing values (one cannot compute log(0)), missing value imputation was performed on TIC normalized data prior to CLR transformation. For the TIC-RCLR transformation, imputation was performed after the transformation, because the missing/0 values are ignored during the process. Similar to the results in Figure 1, there were minimal differences in the total variation explained between the different approaches. PCA for TIC normalization with kNN imputation followed by CLR transformation explained the most variation with the first two principal components (43.02%; Figure 2C), while RF imputation on TIC normalized data followed by CLR transformation captured the least (41.12%; Figure 2D). Clustering of replicates within concentration conditions was also highly similar, and the variation explained using the imputed data was consistently lower than the nonimputed analyses (Figure 1). Similar results were found with transformations performed on the raw data (Figure S2).

To determine the effects of imputation methods on individual MNAR metabolite abundances, we determined the effects of kNN and RF missing-value imputations to non-imputed values on individual metabolites with different levels of missingness. Figures 3A–F show boxplots for a metabolite-level comparison of transformations and imputations where data points are MNAR-type missing. All data points were missing for these spiked-in standards at the two lowest concentrations. TIC-CLR transformation preserved the log-scale proportional relationship between data points in TIC-normalized data, because the data are converted to log-ratios, relative to the geometric mean (Figures 3B and 3C). The TIC-CLR transformation was more similar to the TIC-normalized data, compared with the TIC-RCLR transformation with minimum value imputation (Figure 3D), although both preserved the relationships between the different concentration

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

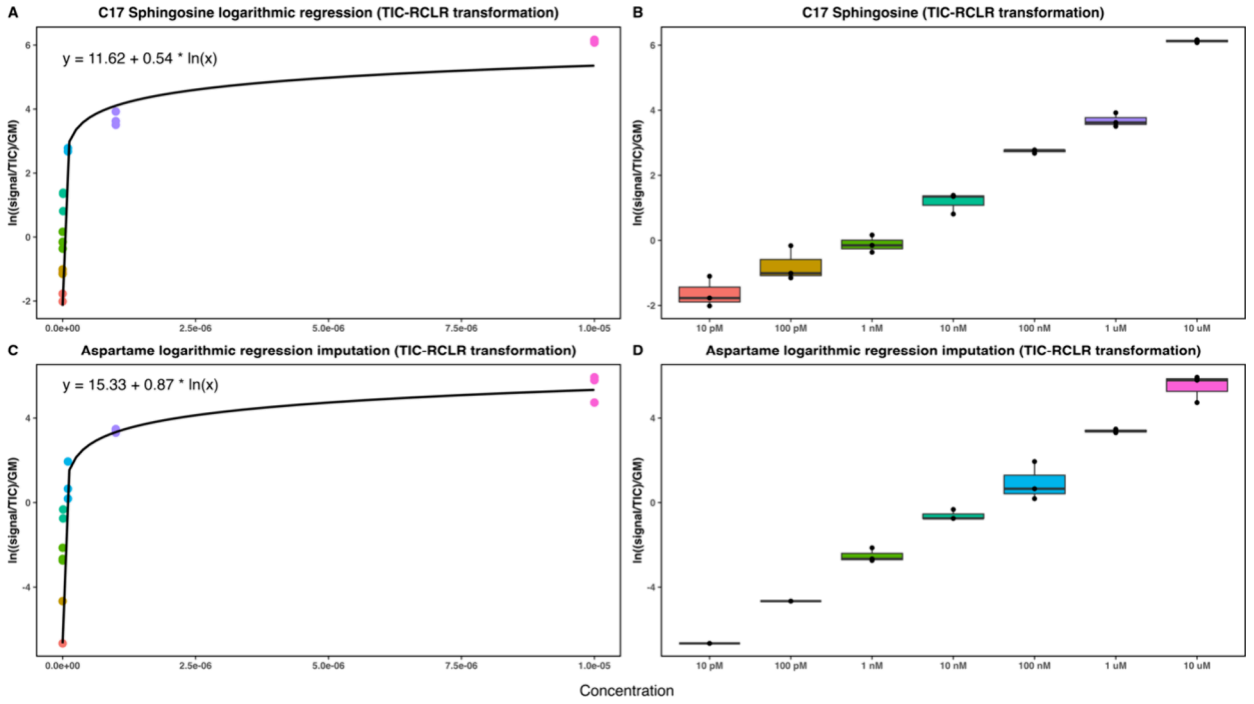


Figure 5. Regression-based modeling approach for handling missing values in targeted metabolomics data which conform to the MNAR-type pattern. (A) Logarithmic regression for a C17 sphingosine, which had detected signals at all concentrations and for all replicates (TIC-RCLR-transformed data). (B) Boxplot for C17 sphingosine. (C) Logarithmic regression and imputation for aspartame, which had missing signals at 2 lowest concentrations (10 and 100 pM) for all replicates (TIC-RCLR transformed data). (D) Boxplot assessing logarithmic regression imputation, following TIC-RCLR transformation for aspartame.

standards. Following the TIC-RCLR transformation, both kNN and RF imputations failed to effectively impute missing values below the limit of detection (Figures 3E and Figure 3F). Both kNN and RF overestimated the missing data, with a larger overestimation using RF compared with kNN imputation (Figure 3F). Nearly identical results were determined when transformations were performed using raw data (Figure S3).

Figures 4A–F show another metabolite-level comparison with MNAR-type missing data. In this example, an additional two orders of magnitude are missing. Similar to Figure 3, the TIC-RCLR transformation followed by kNN and RF imputations failed to effectively impute missing values below the limit of detection (Figures 4E and 4F). RF overestimated missing values more than kNN imputation (Figure 4E and 4F). Nearly identical results were determined when transformations were performed using raw data (Figure S4).

Figures 5A–D show model-based (logistic regression) imputation to handle MNAR-type missing data points. On the log scale, metabolites that had detected signals for all concentrations exhibited a linear relationship (Figure 5B). To impute MNAR missing values, a logistic regression model was fitted using only existing data points and then used to impute missing values (Figure 5C). Following logistic regression model-based imputation, imputed values conform to the expected pattern for data points with order of magnitude concentration differences (Figure 5D).

How Do Compositional Transformations and Imputations Affect Untargeted Metabolomics Data, Which Are Both MCAR and MNAR? The second dataset was a metabolomics analysis of subjects on a vegan or omnivore diet.[46] The dataset had missing values that were both MCAR and MNAR. These MCAR missing values had no discernible pattern to their missingness, which might have been a result of issues with sample handling or detection. We also found missing values that were MNAR because they were present in all of the omnivore diet samples but absent in all of the vegetarian diet samples (and vice versa). We expect many environmental datasets will have some combination of MCAR and MNAR missing values, and our results show that the type of missing data has a profound impact on imputation accuracy.

To assess the impacts of normalization and compositional transformation on MCAR/MNAR data, we compared PCA clustering of samples within the different diets without machine learning imputations (Figures 6A–D). PCA for raw data explained the most variation in the first two principal components (79.92%; Figure 6A), while PCA for the TIC-RCLR transformed data with minimum value imputation captured the least (70.57%; Figure 6D). Clustering of replicates within conditions was similar; four discrete clusters were observed with raw data, TIC normalized data, and TIC-RCLR transformed data with minimum value imputation (Figures 6A, 6B, 6D). PCA for TIC-CLR transformed data failed to resolve the clustering of samples with the same resolution, only resolving the separate diets into two discrete groups (Figure 6C). Nearly identical results were determined when transformations were performed using raw data (Figure S5).

We then determined the effects of kNN and RF imputation methods on PCA clustering and visualization. Figures 7A–D show PCA analyses for the untargeted dataset using machine learning imputations on missing data with TIC-RCLR and TIC-CLR transformations. The total variation captured among the different methods was very similar. PCA for TIC-normalized data with kNN imputation followed by CLR transformation captured the most variation on the first two

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

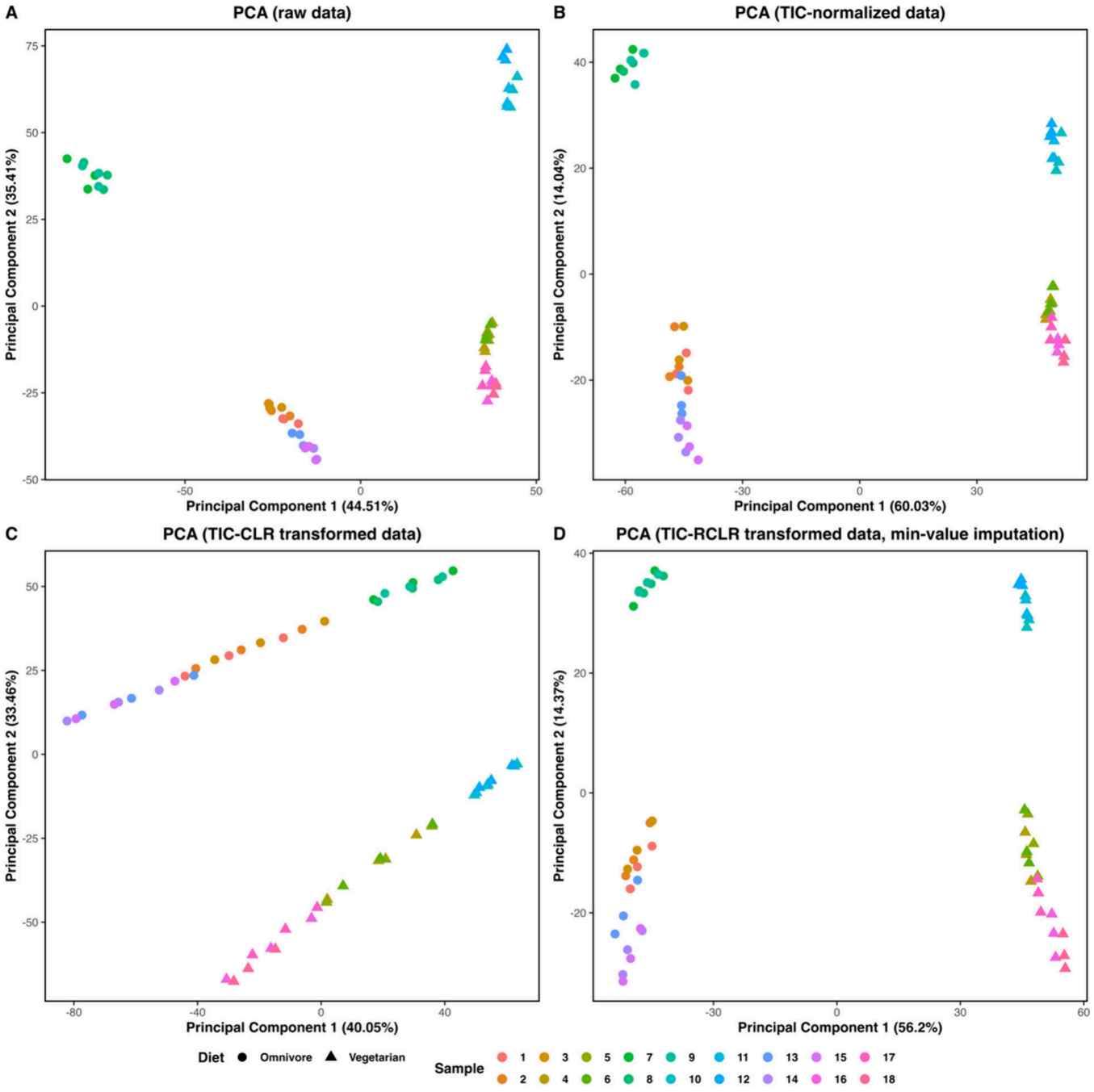


Figure 6. PCA plots for normalizations on untargeted metabolomics data with no machine-learning-based imputations. (A) PCA plot for raw data. (B) PCA plot for TIC-normalized data (pseudocount = 1×10^{-12}). (C) PCA plot for TIC-CLR transformed data. (D) PCA plot for TIC-RCLR transformed data, followed by minimum value imputation for missing values.

principal components (69.77%; Figure 7C), while RF imputation on TIC-normalized data followed by CLR transformation captured the least (68.34%; Figure 7D). Clustering of replicates within concentration conditions was highly similar in all four analyses. While the difference is smaller, we note that the variance explained using the imputed data was still consistently lower than the variance explained using the nonimputed data (Figure 6). Similar results were determined when transformations were performed using raw data (Figure S6).

To test the hypothesis that imputation accuracy increases as the proportion of missing data across features decreases, we binned the percentage of missing data within a feature into three discrete categories: features that had <10% missing values, features that had 50% missing values, and features that had >90% missing values. For each threshold of missing data, we substituted another missing value for an entry that had a known value and assessed whether the accuracy of the ML-based imputation methods improved. Figures 8A–D shows distance comparisons between the true value and imputed value for artificially produced missing values at different thresholds of missing data within a feature for untargeted metabolomics data. Whether missing data is imputed prior to transformation or after, the absolute distance between true

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

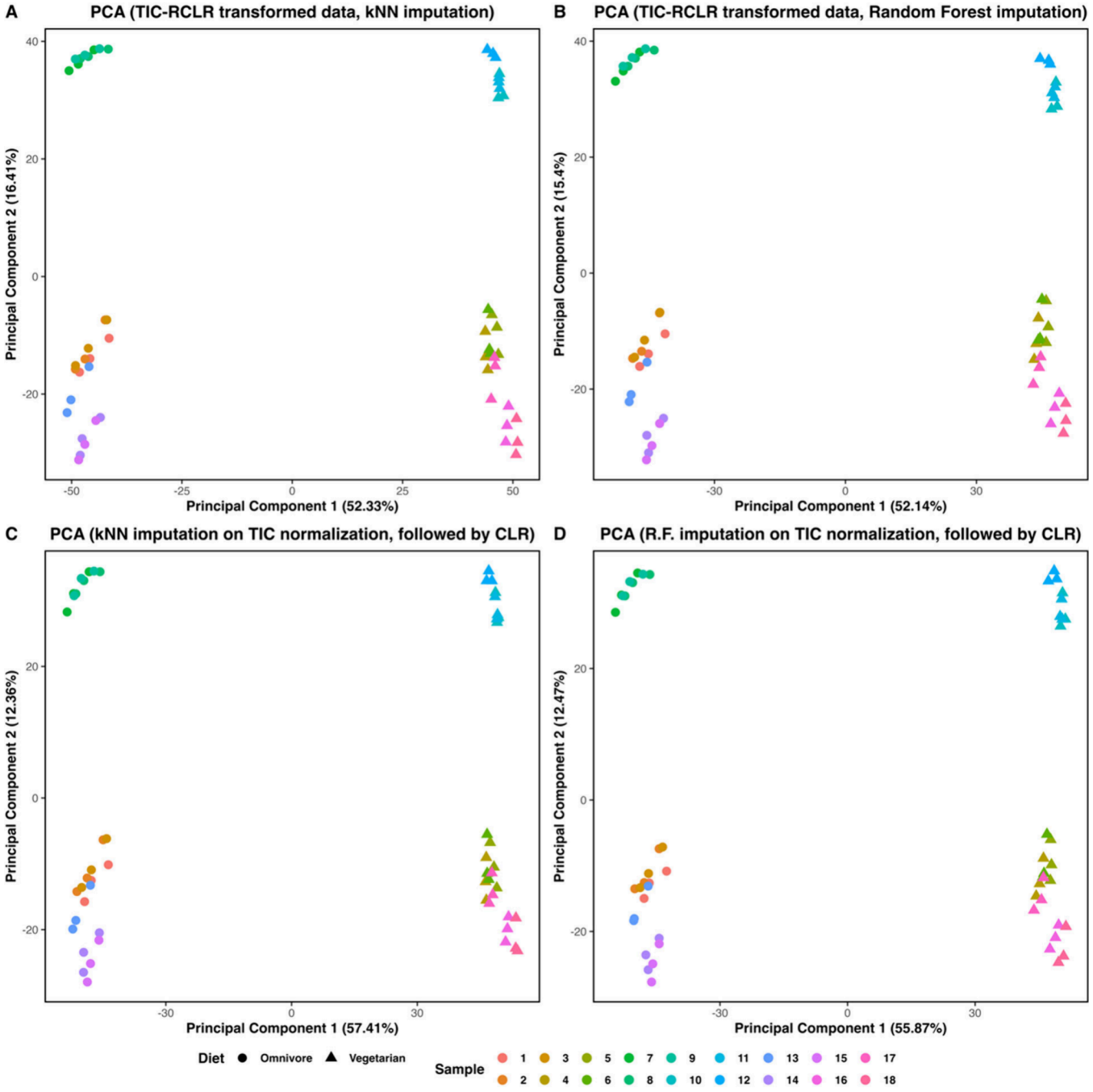


Figure 7. PCA plots of the untargeted metabolomics data using machine learning imputations and two different log-ratio transformation methods. (A) TIC-RCLR transformed data followed by kNN imputation. (B) TIC-RCLR transformed data followed by RF imputation. (C) kNN imputation on TIC normalized data followed by CLR transformation. (D) RF imputation on TIC normalized data, followed by CLR transformation.

value and ML-imputed imputed value varies inversely as the percentage of detected signals increases across samples for a metabolite.

Finally, to determine the effects of the imputation methods on the untargeted MCAR/MNAR metabolite abundances, we performed kNN and RF missing-value imputations on individual metabolites with different proportions of missing data. Figures 9A–F shows a metabolite-level comparison of transformations and imputations where data points are MCAR-type missing within the untargeted dataset using a metabolite with >90% detected signal across all samples. TIC-CLR transformation preserved the proportional relationship between data points observed in TIC-normalized data (Figures 9B and 9C). TIC-RCLR transformation followed by kNN and RF imputations yielded similar results; both methods imputed missing values close to the detected signals (Figures 9E and 9F). Nearly identical results were determined when transformations were performed using raw data (Figure S7).

Figures 10A–F show a metabolite-level comparison of transformations and imputations where data points are MCAR-type missing within the untargeted dataset; the metabolite had 50% detected and 50% undetected signals across all samples, with missing values in both diet groups. TIC-RCLR transformation followed by minimum value imputation preserved

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

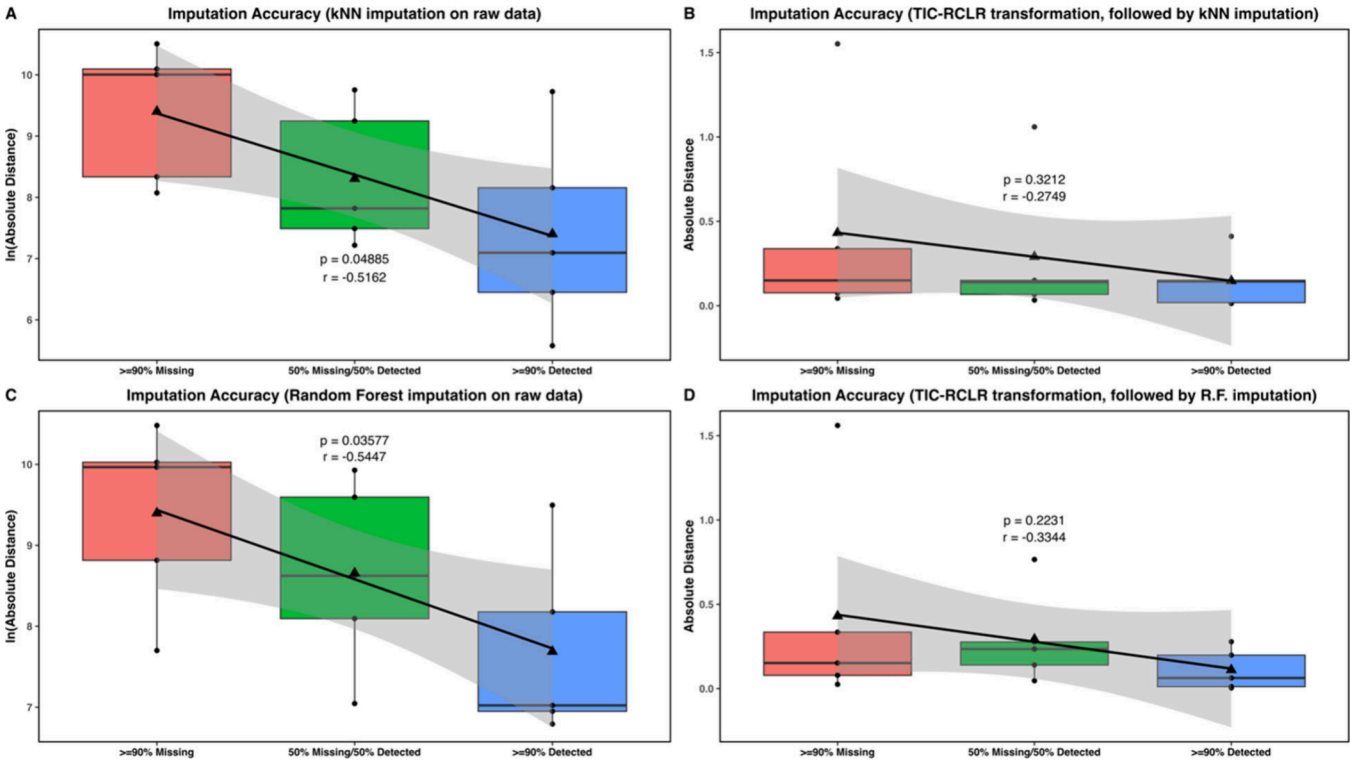


Figure 8. Distance comparisons between true value and imputed value applied to nontransformed untargeted metabolomics data and RCLR transformed data. (A) kNN imputation for missing values on raw data. (B) TIC-RCLR transformation on raw data followed by kNN imputation for missing values. (C) RF imputation for missing values on raw data. (D) TIC-RCLR transformation on raw data, followed by RF imputation for missing values.

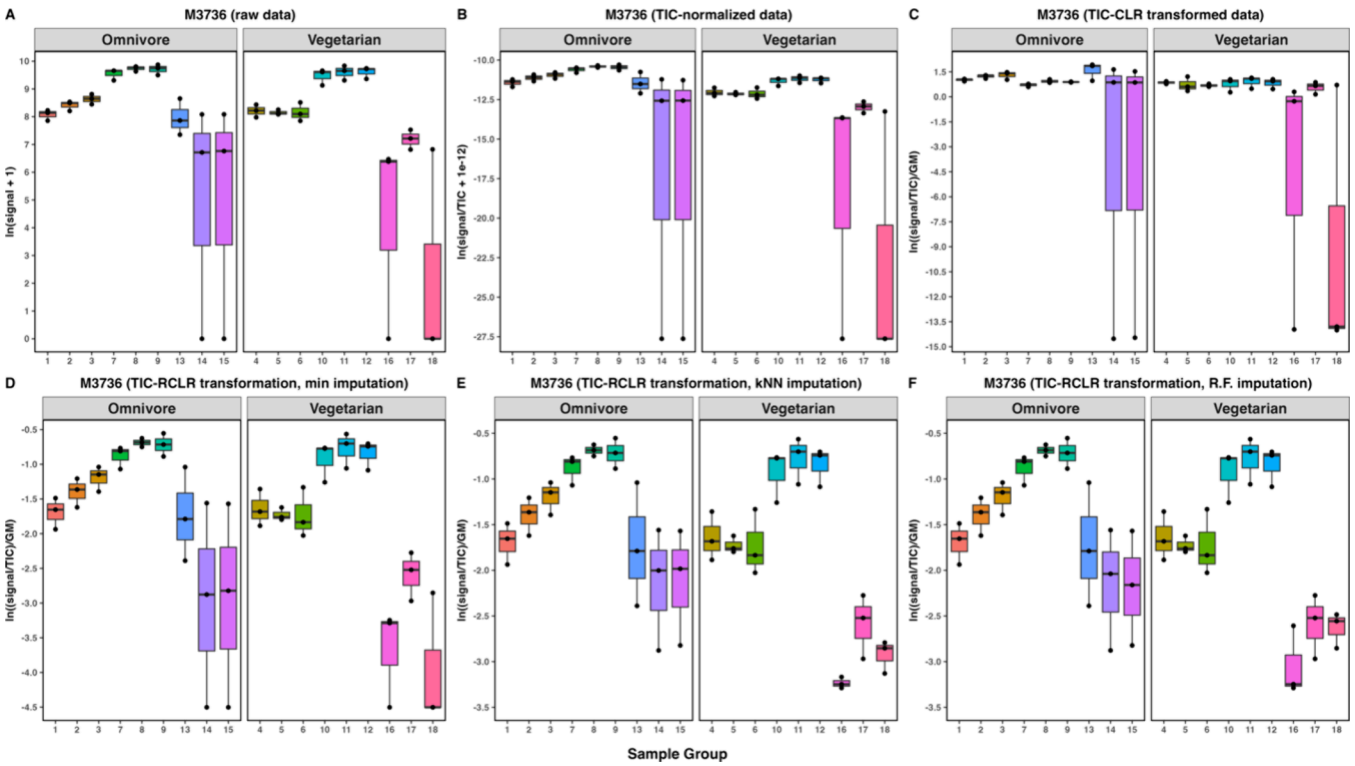


Figure 9. Boxplots of transformations and imputations for a metabolite (ID No. 3736) with >90% detected signals across all samples in the untargeted dataset. (A) Raw data plotted on a natural log scale. (B) TIC normalized data plotted on a natural log scale. (C) TIC-CLR transformed data (pseudocount = 1×10^{-12}). (D) TIC-RCLR transformed data, followed by minimum value imputation. (E) TIC-RCLR transformed data, followed by kNN imputation. (F) TIC-RCLR transformed data, followed by RF imputation.

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

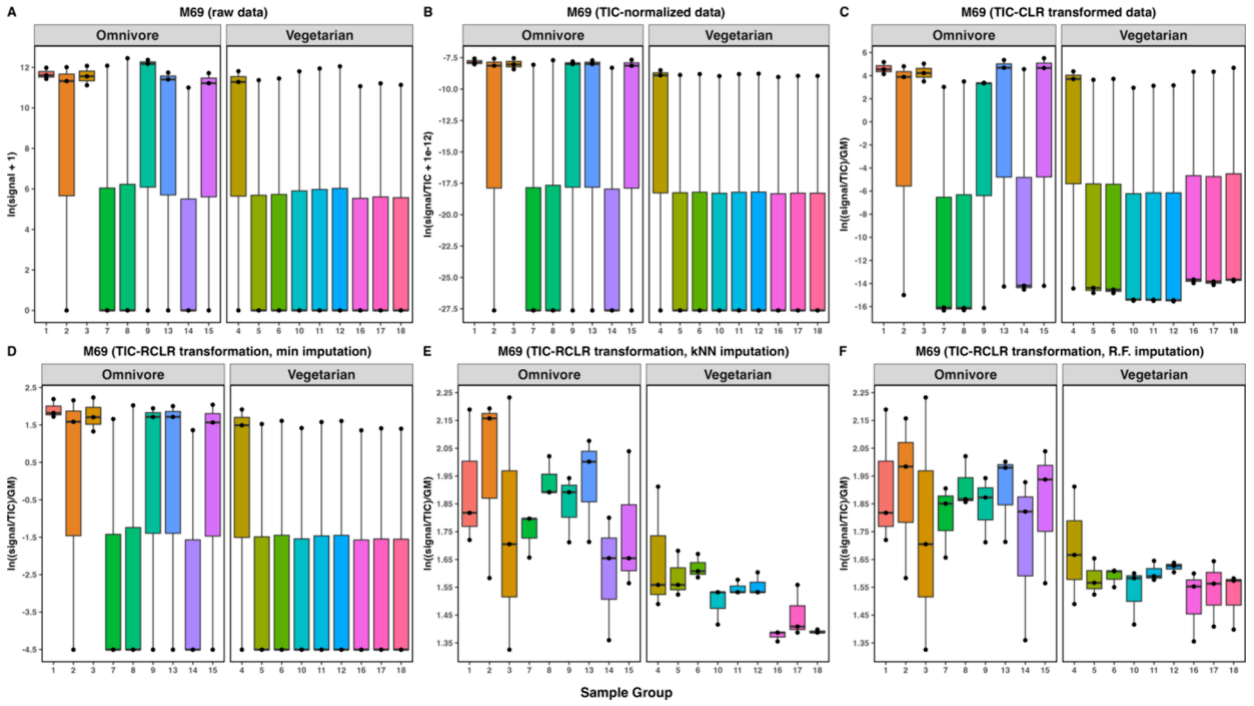


Figure 10. Boxplots of transformations and imputations for a metabolite (ID No. 69) with 50% detected signals, 50% missing signals across all samples in the untargeted dataset. (A) Raw data plotted on a natural log scale. (B) TIC normalized data plotted on a natural log scale. (C) TIC-CLR transformed data (pseudocount = 1×10^{-12}). (D) TIC-RCLR transformed data, followed by minimum value imputation. (E) TIC-RCLR transformed data, followed by kNN imputation. (F) TIC-RCLR transformed data, followed by RF imputation.

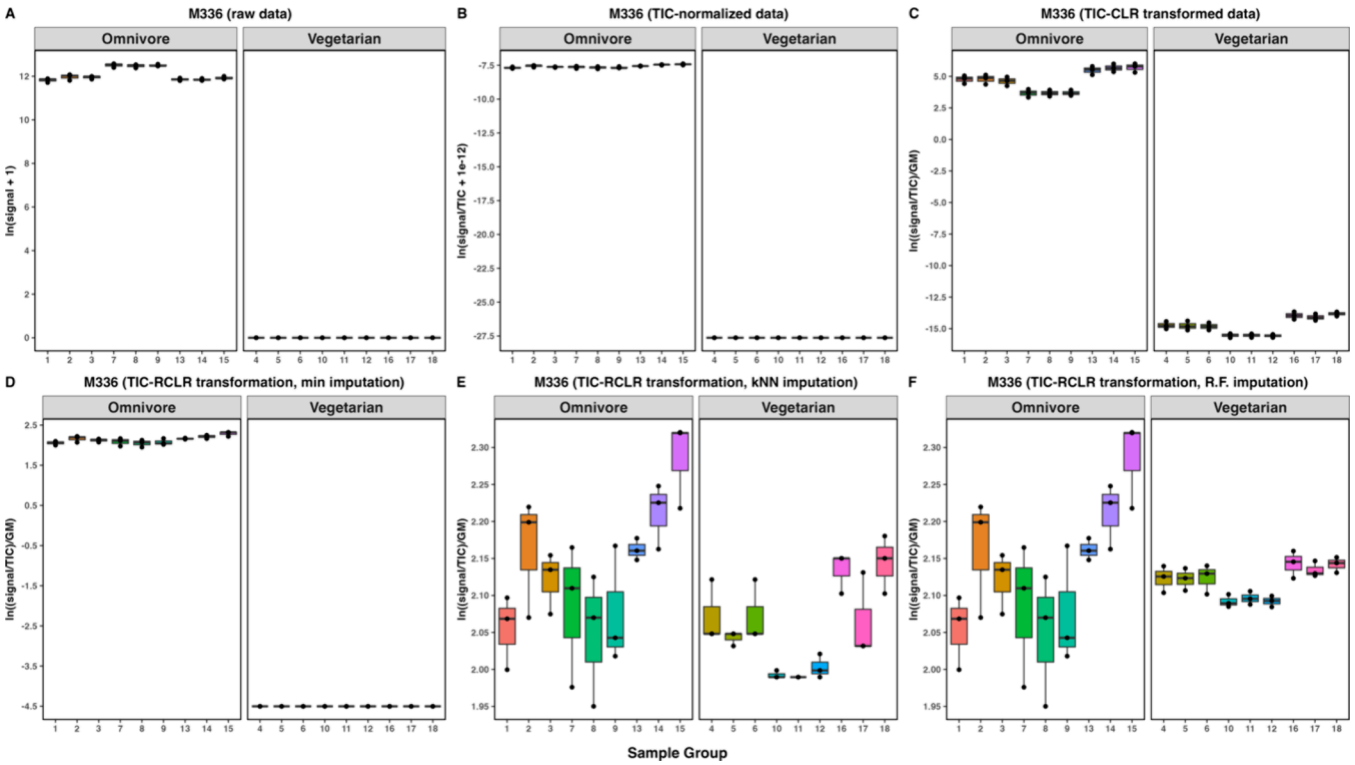


Figure 11. Boxplots of transformations and imputations for a metabolite (ID No. 336) with 50% detected signals, 50% missing signals across all samples in the untargeted dataset. (A) Raw data plotted on a natural log scale. (B) TIC normalized data plotted on a natural log scale. (C) TIC-CLR transformed data (pseudocount = 1×10^{-12}). (D) TIC-RCLR transformed data, followed by minimum value imputation. (E) TIC-RCLR transformed data, followed by kNN imputation. (F) TIC-RCLR transformed data, followed by RF imputation.

the proportional relationship between data points observed in the TIC normalized data (Figures 10B and 10D), while TIC-CLR transformation did not (Figure 10C) due to centering data points around the sample-specific geometric mean and

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

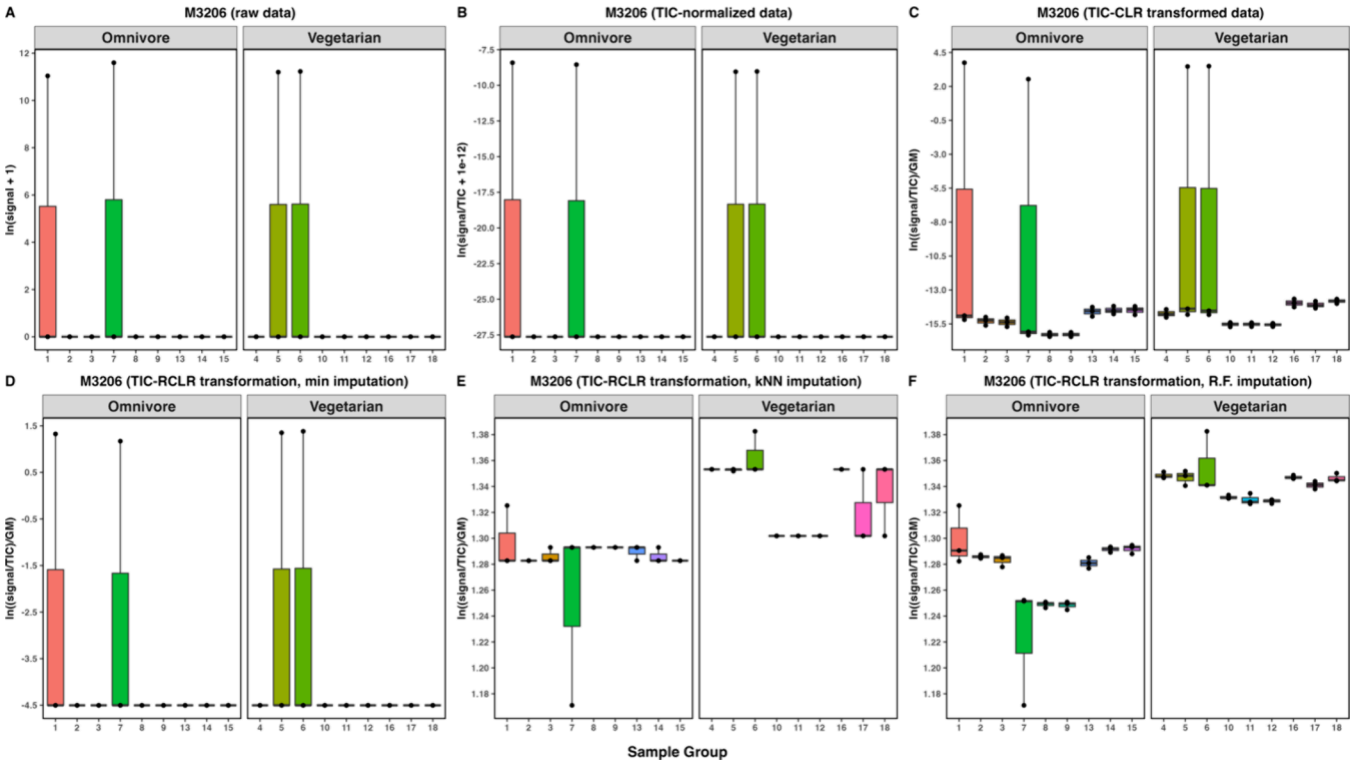


Figure 12. Boxplots of transformations and imputations for a metabolite <10% detected signals across all samples in the untargeted dataset. (A) Raw data plotted on a natural log scale. (B) TIC normalized data plotted on a natural logarithmic scale. (C) TIC-CLR transformed data (pseudocount = 1×10^{-12}). (D) TIC-RCLR transformed data followed by a minimum value imputation. (E) TIC-RCLR transformed data, followed by kNN imputation. (F) TIC-RCLR transformed data, followed by RF imputation.

bias introduced by 0-substitution with pseudocounts during transformation. TIC-RCLR transformation followed by kNN and RF imputation gave similar results; both methods imputed missing values close to the detected signals (Figures 10E and 10F). Nearly identical results were determined when transformations were performed using raw data (Figure S8).

Figures 11A–F show a metabolite-level comparison of transformations and imputations where data points are MNAR-type missing within the untargeted dataset. While the previous metabolites were MCAR, this example showcases MNAR data in the untargeted context, because the metabolite had 50% detected and 50% undetected signals across all samples; however, undetected signals were only in the vegetarian diet samples. Due to the inherent nature of the CLR transformation, it is inevitable to introduce some bias with the transformation attributed to the pseudocount; even with a very small pseudocount (1×10^{-12}), interpretations could lead to misleading results because of division by the geometric mean of each sample during the transformation (Figure 11C). TIC-RCLR transformation followed by minimum value imputation preserved the proportional relationship between data points observed in the TIC-normalized data (Figures 11B and 11D) while TIC-CLR transformation did not (Figure 11C). Both kNN and RF imputations produced completely fabricated results, because of the fact that there was no data to impute from within the same group of individuals (vegetarian)—the only detected signals for the metabolite all came from samples which had an omnivore diet (Figures 11E and 11F). Naively, the imputations give the illusion that the metabolite profile was similar among individuals in both diets. However, the metabolite was completely undetectable in individuals on a vegetarian diet. Nearly identical results were determined when transformations were performed using raw data (Figure S9).

Figures 12A–F show a metabolite-level comparison of transformations and imputations where the metabolite had <10% detected signals across all samples. TIC-RCLR transformation followed by minimum value imputation preserved the proportional relationship between data points observed in the TIC-normalized data (Figures 12B and 12D), while TIC-CLR transformation did not (Figure 12C). TIC-RCLR transformation followed by kNN and RF imputation gave different results; RF imputation consistently had greater imputed values when compared with kNN imputation, a similar finding when compared with the targeted dataset (Figures 12E and 12F). Both RF and kNN imputed missing values, giving the illusion that there were discernible differences between diets; however, this metabolite was almost completely undetectable across all samples. Nearly identical results were determined when transformations were performed using raw data (Figure S10).

Given the clear impact of missing data on imputation accuracy in our unsimulated targeted and untargeted datasets, we directly explored the impacts of different levels of missingness on MCAR imputation accuracy. Table 1 summarizes a comparison of RF- and kNN-based imputations where data are MCAR. RF imputation consistently outperformed kNN imputation, in terms of accuracy assessed by RMSE, concordant with previous publications.[27,47] In both cases, the accuracy declines as the percentage of missing data increases.

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

Table 1. Normalized Root Mean Square Error Assessment of RF and kNN Imputations at Different Thresholds of Randomly Assigned Missing Data for Untargeted Metabolomics

% missing	NRMSE: TIC-RCLR, followed by RF	NRMSE: TIC-RCLR, followed by kNN
20	0.3405	0.3701
40	0.3503	0.3864
60	0.3678	0.4071
80	0.4136	0.5284

DISCUSSION

In this study, we analyzed the effects of imputation on metabolomics data using several widely recognized methods: simple substitution, *k*-Nearest Neighbors, and Random Forest. The primary objective was to understand how these imputation techniques were affected by levels of missing data and by the types of missing data, e.g., MCAR and MNAR. Our findings showed that both kNN and RF imputation methods worked well at low proportions of missing data across samples and when the data were MCAR. We also showed that compositional data methods performed well across datasets, with little evidence of distortion compared to raw and TIC-normalized methods. However, the accuracy of kNN and RF imputations declined markedly as the proportion of missing data increased in both MCAR and MNAR data and both performed especially poorly with MNAR data. At very high levels of missing data and when data are MNAR, kNN and RF imputations falsely imputed differences in metabolite abundance between samples that did not exist in the real data. The effect of overestimation was particularly dramatic with individual metabolites, but the effect was also observed in the PCA analyses of both targeted and untargeted datasets where imputation resulted in a loss of explanatory power. With all methods, it is important to note that imputation often overestimated the abundance of a molecule's signal if it is truly absent from the data, even in the case of minimum value imputation. In such cases, previous studies have recommended stratified imputation (employing prior knowledge to avoid imputation for true zero signals, while imputing values above predefined thresholds), or using combined imputation approaches, employing LOD imputation (minimum-value or instrument LOD), in tandem with multivariate imputations, such as RF.[31]

Our finding that higher levels of missingness in MCAR data corresponded to poorer imputation accuracy makes sense considering that these ML methods train on the existing values in the dataset. The more data available for training, the more accurate is the imputation. In some cases, we discovered that the methods were imputing large numbers of values after training on only a single existing data point. While we only tested the effect of missing data with two of the many dozens of available imputation methods, similar problems likely plague any imputation approach: the fewer the values available for training or estimation, the lower the imputation accuracy. Our discovery that imputation methods performed considerably worse with MNAR also makes sense when we consider that the underlying assumptions of these methods is that the missing data have the same distribution as the nonmissing data. Indeed, the imputation methods were highly accurate when we accurately modeled the MNAR data for the targeted dataset with a logistic regression model. Identifying the pattern of MNAR in the untargeted dataset also showed how imputation assumptions were clearly violated, as they assumed the pattern of missingness was the same for both omnivore and vegetarian samples.

While imputation techniques offer a seemingly convenient solution for handling missing data, they present several significant drawbacks that warrant cautious consideration. The variability inherent in different imputation approaches introduces a degree of uncertainty that can undermine the reliability of the research findings. Imputations are fundamentally based on the existing data. The less data available for imputing, the worse the imputations. As a result, the imputed values might not accurately reflect the true underlying patterns, thereby compromising the integrity of the analysis. This is particularly problematic in cases where the missing data mechanism is not fully understood or is incorrectly assumed to be missing at random (MCAR or MAR) when it is not (MNAR), or vice versa. Our results with real datasets show that the erroneous assumption of the missing data mechanism may lead to inappropriate imputation methods, further exacerbating inaccuracies in the dataset. If one is sure of the classification of missingness, such as with MNAR with log-scale spikes in standards, then an appropriate imputation method can be applied with confidence. However, in many untargeted datasets, the type of missingness may be unclear or even mixed, suggesting great caution should be used with the blind use of imputation methods.

Imputation may work well for data where most of the signals are shared. Such experimental designs might include analysis of organisms where a single gene is deleted, or a single new member is added to a panel of bacteria, or datasets such as plasma metabolites where the same metabolites are observed very frequently (e.g., 50%−100%) across the samples. These study designs tend to reflect the central core metabolism of the samples that are being investigated. However, as factors like diet, microbiome variability, environmental exposures, and medication use differ among individuals, many metabolites may be present infrequently—often in less than 10% of samples within a cohort. Thus, one may have to make choices about whether to perform data imputation. If one filters the data for only common features, as is often done in statistics for untargeted metabolomics data, it is imperative to be transparent and report which data were included and which were excluded. This transparency helps ensure that the analysis remains interpretable and that potential biases are clearly communicated.

CONCLUSION

Imputation methods rely on random processes, model-based predictions, or machine learning techniques. Thus, it is possible that different researchers unknowningly might produce different imputed datasets from the same original data, leading to variability in results. This lack of reproducibility has the potential to undermine the credibility of the research. Depending on the amount of missing data, this could make results difficult to validate or build upon in later work. A prime example of this is demonstrated with machine learning approaches, which typically rely on a pseudorandom number to seed the results (i.e., set.seed() in R and random_state in Python) and be reproducible. Changing these numbers could lead to vastly different results, particularly in the case of RF imputation and other tree-based methods.

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

By introducing artificial data into the analysis, researchers risk presenting findings that are not truly reflective of the real-world phenomena that they aim to study. Therefore, it is advisible to consider alternative strategies for handling missing data, such as robust statistical methods that can accommodate missing data without the need for imputation, where possible designing studies that minimize the occurrence of missing data or have increased number of biological/technical replicates, and compositional data transformations. In light of these considerations, it is prudent to approach the use of imputations with great caution.[48] Ensuring the validity, reliability, and reproducibility of research should take precedence, guiding researchers toward more transparent and scientifically sound methods for managing missing data.

METHODS

Tandem Mass Spectrometry Analysis. We used two publicly available untargeted metabolomics datasets in our analyses. The details of data collection for each dataset are presented below.

Targeted Dataset. Dataset 1 was collected as described in thew work of Melnik et al.[49] The dataset contains a mixture of 41 standards spiked in at equimolar concentrations to 10 uM fecal extract. Standards were added across a concentration gradient of 10, 100, 1, 10, 100, 1, and 10 uM with three replicates per concentration. Briefly, MS/MS data was acquired in positive mode on a Q Exactive Orbitrap (Thermo Fisher Scientific, Waltham, MA) using data-dependent acquisition. Samples were separated using a Vanquish UPLC instrument (Thermo Fisher Scientific, Waltham, MA) on a 100 × 2.1 mm Kinetix 1.7 μM C18 column (Phenomenex, Torrance, CA) using the following buffers. Buffer A was water (J.T. Baker, LC-MS grade) with 0.1% formic acid (Thermo Fisher Scientific, Optima LC/MS). Buffer B was acetonitrile (J.T. Baker, LC-MS grade) with 0.1% formic acid (Fisher Scientific, Optima LC/MS). Flow rate was set to 0.5 mL/min with a gradient of 0−1 min 5% B, 1−8 min 100% B, 8−10.9 min 100% B, 10.9−11 min 5% A, and 11−12 min 5% B. Mass range was set to 100−1500 m/z, MS1 scan level resolution was set to 35K and MS2 scan resolution was set to 17.5K. Data are deposited in MassIVE (www.massive.ucsd.edu) under accession MSV000079760. The targeted dataset contained 21 samples, 1925 metabolic features (33 of which were internal standards), with 6551 (16.2%) missing data points (0, no detected signal).

Untargeted Dataset. Dataset 2 was collected from NIST reference grade test materials (RGTM) containing homogenized fecal matter from subjects with vegan or omnivore diets (RGTM 10162, 10171, 10172, and 10173). Samples were collected from 18 individuals (9 vegan, 9 omnivores) with three technical replicates per individual. MS/MS data was acquired in positive mode on a Q Exactive Orbitrap (Thermo Fisher Scientific, Waltham, MA) using data-dependent acquisition. The untargeted dataset contained 54 samples and 42,44 metabolites, with 32,675 (14.3%) missing data points (0, no detected signal). Data are deposited in MassIVE (www.massive.ucsd.edu), under accession MSV000086989.

Data Processing. Data processing was performed using the GNPS analysis ecosystem and MZmine3 (3.9.0).[50] Raw data were first converted to .mzML, using Proteowizard MSconvert (3.0.22287−170037b), before performing feature finding in MZMine3.[51] All feature finding parameters are included in the Supporting Information. Library annotation and molecular networking were performed using GNPS for Dataset 1 and can be accessed via the following link: (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=94d795a6bfeb411a8f36a43b12b95eea)

Normalizations and Transformations. *Total Ion Count Normalization.* TIC normalization (TICN) was carried out by summation of all signals for each metabolite (s_i)—Total Ion Count (TIC)—within a sample (s) and dividing each metabolite's signal strength by the resulting sum.

$$\mathrm{TIC}(s) = \sum_{i=1}^{n} s_i$$

$$\mathrm{TICN}(s_i) = \frac{s_i}{\mathrm{TIC}(s)}$$

$$\mathrm{TICN}(s) = \frac{s_1}{\mathrm{TIC}(s)}, \frac{s_2}{\mathrm{TIC}(s)}, \ldots, \frac{s_n}{\mathrm{TIC}(s)}$$

Centered Log-Ratio Transformation. The centered log-ratio (CLR) transformation transforms compositional data into real number space by taking the logarithm of each sample's component value and dividing by the geometric mean (GM) of all components, thereby eliminating the constant-sum constraint and making the data suitable for multivariate analysis. The generalized formula for the CLR transformation is shown below where s is a sample vector, s_i is a component (metabolite signal) of the sample vector, CLR(s_i) is an individual log ratio for a component of the vector, and CLR(s) is the transformed vector:

$$\mathrm{GM}(s) = \sqrt[n]{\prod_{i=1}^{n} s_i}$$

$$\mathrm{CLR}(s_i) = \log\left(\frac{s_i}{\mathrm{GM}(s)}\right)$$

$$\mathrm{CLR}(s) = \log\left(\frac{s_1}{\mathrm{GM}(s)}\right), \log\left(\frac{s_2}{\mathrm{GM}(s)}\right), \ldots, \log\left(\frac{s_n}{\mathrm{GM}(s)}\right)$$

In the context of the metabolomics datasets, the raw metabolite signals are not inherently compositional. Before performing the CLR transformation for the datasets, each sample was first TIC-normalized to convert it to compositional form; then the CLR transformation was applied to each sample vector to yield a final TIC-CLR transformation. CLR calculations were performed using the decostand(x, method = "clr", pseudocount = 1×10^{-12}) function within the vegan R library.[52]

Robust Centered Log-Ratio Transformation. The robust centered log ratio (RCLR) transformation is an extension of the traditional CLR transformation, which allows for the presence of zeros within datasets. The RCLR transformation ignores zero values and divides all nonzero values by the geometric mean of the observed features, followed by a log transformation on the nonzero log ratios; the zero values remain unchanged. After the transformation, the unchanged zero values must be handled. The RCLR transformation is calculated similarly to the CLR transformation, with the caveat that 0 and missing signals are not included in the calculation:

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

$$\mathrm{GM}(s) = \sqrt[n]{\prod_{i=1}^{n} s_i} \qquad \text{where } s_i \neq 0$$

$$\mathrm{RCLR}(s_i) = \log\left(\frac{s_i}{\mathrm{GM}(s)}\right)$$

$$\mathrm{RCLR}(s) = \log\left(\frac{s_1}{\mathrm{GM}(s)}\right), \log\left(\frac{s_2}{\mathrm{GM}(s)}\right), ..., \log\left(\frac{s_n}{\mathrm{GM}(s)}\right)$$

Similar to the CLR transformation, each sample was first TIC-normalized to convert it to compositional form, and then the RCLR transformation was applied to each sample vector to yield a final TIC-RCLR transformation. RCLR calculations were performed using the decostand(x, method = "rclr") function within the vegan R library.

Proof That Compositional Transformation on TIC-Normalized Data Removes Compositionality. Letting the vector of data be denoted as $\hat{x}$, we consider be defined as follows:

$$\hat{x} = \{x_1, x_2, ..., x_m\}$$

We say that $\hat{x}$ is a vector in $\mathbb{R}^m$ and denote the ith entry of $\hat{x}$ as x_i. We define the TIC-normalized data vector $\hat{x}^N$ by its ith entry as

$$x_i^N = \frac{x_i}{\sum_{i=1}^{m} x_i}$$

Each entry is the corresponding entry in the non-normalized vector divided by the sum of all entries. This normalization can be shown to provide compositional data because the sum of the elements of the normalized vector, $S(\hat{x}^N)$, can be written as

$$S(\hat{x}^N) = \frac{x_1}{\sum_{i=1}^{m} x_i} + \frac{x_2}{\sum_{i=1}^{m} x_i} + \cdots + \frac{x_m}{\sum_{i=1}^{m} x_i}$$

$$S(\hat{x}^N) = \frac{\sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} x_i} = 1$$

We define the CLR transformation of the TIC-normalized vector by its ith entry as

$$\mathrm{CLR}(x_i^N) = \log\left(\frac{x_i^N}{\sqrt[m]{\prod_{i=1}^{m} x_i^N}}\right)$$

Similarly, the CLR transformation of the un-normalized data vector by its ith entry is defined as

$$\mathrm{CLR}(x_i) = \log\left(\frac{x_i}{\sqrt[m]{\prod_{i=1}^{m} x_i}}\right)$$

We can expand the CLR transform of the normalized vector in the following manner:

$$\mathrm{CLR}(x_i^N) = \log\left(\frac{x_i^N}{\sqrt[m]{\prod_{i=1}^{m} x_i^N}}\right)$$

$$\mathrm{CLR}(x_i^N) = \log\left(\frac{\frac{x_i^N}{\sum_{i=1}^{m} x_i^N}}{\sqrt[m]{\prod_{i=1}^{m} x_i^N}}\right)$$

$$\mathrm{CLR}(x_i^N) = \log\left(\frac{x_i^N}{\sum_{i=1}^{m} x_i^N \cdot \sqrt[m]{\prod_{i=1}^{m} x_i^N}}\right)$$

$$\mathrm{CLR}(x_i^N) = \log(x_i^N) - \log\left(\sum_{i=1}^{m} x_i^N \cdot \sqrt[m]{\prod_{i=1}^{m} x_i^N}\right)$$

$$\mathrm{CLR}(x_i^N) = \log(x_i^N) - \log\left(\sum_{i=1}^{m} x_i^N\right) - \log\left(\sqrt[m]{\prod_{i=1}^{m} x_i^N}\right)$$

A similar expansion results in the ith entry of the CLR transform of the un-normalized vector being written as follows:

$$\mathrm{CLR}(x_i) = \log(x_i) - \log\left(\sqrt[m]{\prod_{i=1}^{m} x_i}\right)$$

We can see that the two different treatments of the data vector only differ by the constant term $\log(\sum_{i=1}^{m} x_i^N)$ and are notably similar in terms of interpretation as noted in the main body text. We then observe the sum of the TIC-CLR transformed data vector to determine if the compositionality has been removed:

$$\sum_{x=i}^{m} \mathrm{CLR}(x_i^N) = \sum_{x=i}^{m} \left(\log(x_i^N) - \log\left(\sum_{i=1}^{m} x_i^N\right) - \log\left(\sqrt[m]{\prod_{i=1}^{m} x_i^N}\right)\right)$$

$$\sum_{x=i}^{m} \mathrm{CLR}(x_i^N) = \sum_{x=i}^{m} \log(x_i^N) - \sum_{x=i}^{m} \log\left(\sum_{i=1}^{m} x_i^N\right) - \sum_{x=i}^{m} \log\left(\sqrt[m]{\prod_{i=1}^{m} x_i^N}\right)$$

$$\sum_{x=i}^{m} \mathrm{CLR}(x_i^N) = \log\left(\prod_{i=1}^{m} x_i^N\right) - m \log\left(\sum_{i=1}^{m} x_i^N\right) - \log\left(\prod_{i=1}^{m} x_i^N\right)$$

$$\sum_{x=i}^{m} \mathrm{CLR}(x_i^N) = -m \log\left(\sum_{i=1}^{m} x_i^N\right)$$

We find that the resulting sum is equal to the negative logarithm of the sum of entries of the original vector multiplied by the size of the vector, thereby showing that the entries do not sum to a fixed value that is not dependent on the data and, thus, the data lacks compositionality.

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

Minimum Value Imputations. Minimum value imputations were carried out for RCLR transformations, substituting nontransformed 0/missing signal values with the minimum value of the TIC-RCLR transformed feature table.

k-Nearest Neighbors Imputations. The *k*-Nearest Neighbors imputations were carried out with the R package VIM,[53] using the kNN function. The kNN function uses the Gower Distance to calculate similarity between samples/variables, as the method allows for mixed data types. In the case of purely quantitative data types (used in this study), the Gower Distance calculations employ only the Manhattan Distance. The kNN method will raise an error if any features (metabolites) contain entirely missing values. The method arguments used were $k = 3$ (three nearest neighbors), numFun = mean (averages the value of the three nearest neighbors), useImputedDist = FALSE (does not use imputed values for subsequent distance calculations), and default values for all other arguments.

Random Forests Imputations. The Random Forests imputations were generated using the R package missForest,[54] with the self-named function missForest(). The parameters used were 10 for *maxiter* (maximum number of iterations if convergence has not yet been reached), 100 for *ntree* (number of trees to grow in each forest), and default values for all other arguments.

The general steps to the iterative algorithm are (imputations are made in order for samples/variables with the least amount of missing values, moving toward samples/variables with more missing values):

(1) Initialize the imputation process by making initial guesses for missing values with the mean of the matrix.
(2) For each sample/variable with missing values, a random forest is trained using other samples/variables as predictors; the training is done only on observed values.
(3) The trained random forests model predicts the missing values for the sample/variable.
(4) The missing values for the sample/variable are updated with each iteration.
(5) The process is repeated iteratively until the imputed values reach convergence (or a predefined number of iterations is specified).
(6) After convergence (or maximum iterations), the imputed dataset is returned.

Details of the full algorithm are specified in the source publication.

Targeted Metabolomics Dataset Methods. To assess the effects of machine learning imputations and model-based imputation when the missing mechanism is MNAR (data are missing systematically at the lowest concentration/at the limit of detection), the targeted dataset was utilized. This dataset contained 41 internal standard metabolites which were spiked-in at fixed concentrations of 10 pM, 100 pM, 1 nM, 10 nM, and 100 nM. For machine learning imputations, all missing values were imputed with both kNN and RF with the parameters discussed above. For modeling-based imputations, a logistic regression was fit to the detected data and then used to predict values for missing data.

Untargeted Metabolomics Dataset Methods. The effectiveness of machine-learning-based imputations for MCAR-type and MNAR-type data was assessed using the untargeted dataset. To directly compare the performance of RF and kNN imputations, the dataset was TIC-RCLR transformed, then missing values were randomly assigned at 20%, 40%, 60%, and 80% intervals using the prodNA() function within the missForest package. Each dataset with missing values at the different intervals was imputed separately, then imputation performance was assessed by normalized root-mean-square error (NRMSE) with the missForest function nrmse(), which accepts the imputed dataset, the dataset with missing (NA) values, and the true dataset (TIC-RCLR transformed, no missing/imputed values) as arguments. Metabolite-level imputation accuracy was also assessed by absolute distance between true value and imputed value at three discrete intervals: metabolites which had ≤10% detected signals across all samples, metabolites with 50% detected signals across all samples, and metabolites which had ≥90% detected signals across all samples. Correlation was determined with cor.test(method = "pearson").

■ ASSOCIATED CONTENT

Data Availability Statement

The targeted metabolomics dataset used for analysis is deposited in MassIVE (massive.ucsd.edu), under accession MSV000079760. The untargeted metabolomics dataset used for analysis is deposited in MassIVE, under accession MSV000086989. The R Markdown scripts used for analyses and visualizations are available at https://github.com/ddkrutkin/metabolomics_imputations

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jasms.4c00434.

Supplemental figures S1−S10 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Scott T. Kelley − *Department of Biology, San Diego State University, San Diego, California 92182, United States;* orcid.org/0000-0001-9547-4169; Email: skelley@sdsu.edu

Authors

Dennis D. Krutkin − *School of Biological Sciences, University of California San Diego, La Jolla, California 92037, United States; Department of Biology, San Diego State University, San Diego, California 92182, United States*

Sydney Thomas − *Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92037, United States*

Simone Zuffa − *Skaggs School of Pharmacy and Pharmaceutical Sciences and Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, California 92037, United States;* orcid.org/0000-0001-7237-3402

Prajit Rajkumar − *Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, California 92037, United States;* orcid.org/0009-0002-5538-5270

Rob Knight − *Department of Computer Science and Engineering, Department of Pediatrics, and Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, California 92037, United States*

Pieter C. Dorrestein − *Skaggs School of Pharmacy and Pharmaceutical Sciences, Collaborative Mass Spectrometry Innovation Center, Department of Pediatrics, and Center for*

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX

Microbiome Innovation, University of California San Diego, La Jolla, California 92037, United States

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.4c00434

Notes

The authors declare the following competing financial interest(s): P.C.D. is an advisor and holds equity in Cybele, BileOmix and Sirenas and a Scientific cofounder, advisor and holds equity to Ometa, Enveda, and Arome with prior approval by UC-San Diego. P.C.D. also consulted for DSM Animal Health in 2023. R.K. is a scientific advisory board member, and consultant for BiomeSense, Inc., has equity and receives income. R.K. is a scientific advisory board member and has equity in GenCirq. R.K. is a consultant and scientific advisory board member for DayTwo and receives income. R.K. has equity in and acts as a consultant for Cybele. R.K. is a cofounder of Biota, Inc., and has equity. R.K. is a cofounder and a scientific advisory board member of Micronoma and has equity. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict-of-interest policies. S.P.T. is employed by Ometa Laboratories as an Applications Scientist.

■ ACKNOWLEDGMENTS

Dennis D. Krutkin's graduate studies are supported by a Division of Research and Innovation Doctoral Fellowship.

■ REFERENCES

(1) Mohanty, I.; Mannochio-Russo, H.; Schweer, J. V.; El Abiead, Y.; Bittremieux, W.; Xing, S.; Schmid, R.; Zuffa, S.; Vasquez, F.; Muti, V. B.; et al. The Underappreciated Diversity of Bile Acid Modifications. *Cell* **2024**, *187* (7), 1801−1818.e20 e20.

(2) Dilmore, A. H.; Martino, C.; Neth, B. J.; West, K. A.; Zemlin, J.; Rahman, G.; Panitchpakdi, M.; Meehan, M. J.; Weldon, K. C.; Blach, C.; et al. Effects of a Ketogenic and Low-Fat Diet on the Human Metabolome, Microbiome, and Foodome in Adults at Risk for Alzheimer's Disease. *Alzheimer's Dement.* **2023**, *19* (11), 4805−4816.

(3) Richards, A. L.; Eckhardt, M.; Krogan, N. J. Mass Spectrometry-based Protein-Protein Interaction Networks for the Study of Human Diseases. *Mol. Syst. Biol.* **2021**, *17* (1), No. e8792.

(4) Roje, B.; Zhang, B.; Mastrorilli, E.; Kovačić, A.; Sušak, L.; Ljubenkov, I.; Ćosić, E.; Vilović, K.; Meštrović, A.; Vukovac, E. L.; Bučević-Popović, V.; Puljiz, Ž.; Karaman, I.; Terzić, J.; Zimmermann, M. Gut Microbiota Carcinogen Metabolism Causes Distal Tissue Tumours. *Nature* **2024**, *632* (8027), 1137−1144.

(5) Fogelson, K. A.; Dorrestein, P. C.; Zarrinpar, A.; Knight, R. The Gut Microbial Bile Acid Modulation and Its Relevance to Digestive Health and Diseases. *Gastroenterology* **2023**, *164* (7), 1069−1085.

(6) Maruvada, P.; Lampe, J. W.; Wishart, D. S.; Barupal, D.; Chester, D. N.; Dodd, D.; Djoumbou-Feunang, Y.; Dorrestein, P. C.; Dragsted, L. O.; Draper, J.; Duffy, L. C.; Dwyer, J. T.; Emenaker, N. J.; Fiehn, O.; Gerszten, R. E.; B Hu, F.; Karp, R. W.; Klurfeld, D. M.; Laughlin, M. R.; Little, A. R.; Lynch, C. J.; Moore, S. C.; Nicastro, H. L.; O'Brien, D. M.; Ordovás, J. M.; Osganian, S. K.; Playdon, M.; Prentice, R.; Raftery, D.; Reisdorph, N.; Roche, H. M.; Ross, S. A.; Sang, S.; Scalbert, A.; Srinivas, P. R.; Zeisel, S. H. Perspective: Dietary Biomarkers of Intake and Exposure—Exploration with Omics Approaches. *Adv. Nutr.* **2020**, *11* (2), 200−215.

(7) Voronov, G.; Lightheart, R.; Frandsen, A.; Bargh, B.; Haynes, S. E.; Spencer, E.; Schoenhardt, K. E.; Davidson, C.; Schaum, A.; Macherla, V. R.; DeBloois, E.; Healey, D.; Kind, T.; Dorrestein, P.; Colluru, V.; Butler, T.; Yu, M. S. MS2Prop: A Machine Learning Model That Directly Generates de Novo Predictions of Drug-Likeness of Natural Products from Unannotated MS/MS Spectra. bioRxiv May 30, *bioRXiv***2024**; p 2022.10.09.511482. .

(8) Webb-Robertson, B.-J. M.; Wiberg, H. K.; Matzke, M. M.; Brown, J. N.; Wang, J.; McDermott, J. E.; Smith, R. D.; Rodland, K. D.; Metz, T. O.; Pounds, J. G.; Waters, K. M. Review, Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-Based Label-Free Global Proteomics. *J. Proteome Res.* **2015**, *14* (5), 1993−2001.

(9) Hrydziuszko, O.; Viant, M. R. Missing Values in Mass Spectrometry Based Metabolomics: An Undervalued Step in the Data Processing Pipeline. *Metabolomics* **2012**, *8* (1), 161−174.

(10) Fraier, D.; Ferrari, L.; Heinig, K.; Zwanziger, E. Inconsistent Internal Standard Response in LC-MS/MS Bioanalysis: An Evaluation of Case Studies. *Bioanalysis* **2019**, *11* (18), 1657−1667.

(11) Buonarati, M. H.; Schoener, D. Investigations Beyond Standard Operating Procedure on Internal Standard Response. *Bioanalysis* **2019**, *11* (18), 1669−1678.

(12) Piehowski, P. D.; Petyuk, V. A.; Orton, D. J.; Xie, F.; Moore, R. J.; Ramirez-Restrepo, M.; Engel, A.; Lieberman, A. P.; Albin, R. L.; Camp, D. G.; Smith, R. D.; Myers, A. J. Sources of Technical Variability in Quantitative LC-MS Proteomics: Human Brain Tissue Sample Analysis. *J. Proteome Res.* **2013**, *12* (5), 2128−2137.

(13) Li, W.; Zhang, J.; Tse, F. L. S. Strategies in Quantitative LC-MS/MS Analysis of Unstable Small Molecules in Biological Matrices. *Biomed. Chromatogr.* **2011**, *25* (1−2), 258−277.

(14) Panda, D.; Dash, B. P.; Manickam, S.; Boczkaj, G. Recent Advancements in LC-MS Based Analysis of Biotoxins: Present and Future Challenges. *Mass Spectrom. Rev.* **2022**, *41* (5), 766−803.

(15) Aszyk, J.; Byliński, H.; Namieśnik, J.; Kot-Wasik, A. Main Strategies, Analytical Trends and Challenges in LC-MS and Ambient Mass Spectrometry-Based Metabolomics. *TrAC Trends Anal. Chem.* **2018**, *108*, 278−295.

(16) Hrydziuszko, O.; Viant, M. R. Missing Values in Mass Spectrometry Based Metabolomics: An Undervalued Step in the Data Processing Pipeline. *Metabolomics* **2012**, *8* (1), 161−174.

(17) Cambiaghi, A.; Ferrario, M.; Masseroli, M. Analysis of Metabolomic Data: Tools, Current Strategies and Future Challenges for Omics Data Integration. *Brief. Bioinform.* **2017**, *18* (3), 498−510.

(18) de Souto, M. C.; Jaskowiak, P. A.; Costa, I. G. Impact of Missing Data Imputation Methods on Gene Expression Clustering and Classification. *BMC Bioinformatics* **2015**, *16* (1), 64.

(19) 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* (1), 663.

(20) Kokla, M.; Virtanen, J.; Kolehmainen, M.; Paananen, J.; Hanhineva, K. Random Forest-Based Imputation Outperforms Other Methods for Imputing LC-MS Metabolomics Data: A Comparative Study. *BMC Bioinformatics* **2019**, *20* (1), 492.

(21) Woods, A. D.; Gerasimova, D.; Van Dusen, B.; Nissen, J.; Bainter, S.; Uzdavines, A.; Davis-Kean, P. E.; Halvorson, M.; King, K. M.; Logan, J. A. R.; Xu, M.; Vasilev, M. R.; Clay, J. M.; Moreau, D.; Joyal-Desmarais, K.; Cruz, R. A.; Brown, D. M. Y.; Schmidt, K.; Elsherif, M. M. Best Practices for Addressing Missing Data through Multiple Imputation. *Infant Child Dev.* **2024**, *33* (1), No. e2407.

(22) Enders, E. *Applied Missing Data Analysis*, Second ed..; Guilford Publications.

(23) Mainzer, R.; Moreno-Betancur, M.; Nguyen, C.; Simpson, J.; Carlin, J.; Lee, K. Handling of Missing Data with Multiple Imputation in Observational Studies That Address Causal Questions: Protocol for a Scoping Review. *BMJ. Open* **2023**, *13* (2), No. e065576.

(24) Mante, J.; Gangadharan, N.; Sewell, D. J.; Turner, R.; Field, R.; Oliver, S. G.; Slater, N.; Dikicioglu, D. A Heuristic Approach to Handling Missing Data in Biologics Manufacturing Databases. *Bioprocess Biosyst. Eng.* **2019**, *42* (4), 657−663.

(25) Zhang, Z. Missing Data Imputation: Focusing on Single Imputation. *Ann. Transl. Med.* **2016**, *4* (1), 9.

(26) Taylor, S. L.; Ruhaak, L. R.; Kelly, K.; Weiss, R. H.; Kim, K. Effects of Imputation on Correlation: Implications for Analysis of Mass Spectrometry Data from Multiple Biological Matrices. *Brief. Bioinform.* **2016**, *18* (2), 312−320.

(27) Gromski, P. S.; Xu, Y.; Kotze, H. L.; Correa, E.; Ellis, D. I.; Armitage, E. G.; Turner, M. L.; Goodacre, R. Influence of Missing Values Substitutes on Multivariate Analysis of Metabolomics Data. *Metabolites* **2014**, *4* (2), 433−452.
(28) 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* (1), 663.
(29) *proportion of missing data should not be used to guide decisions on multiple imputation - ScienceDirect*. https://www.sciencedirect.com/science/article/pii/S0895435618308710 (accessed 2024−10−25).
(30) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. MetaboAnalyst: A Web Server for Metabolomic Data Analysis and Interpretation. *Nucleic Acids Res.* **2009**, *37*, W652−W660.
(31) Do, K. T.; Wahl, S.; Raffler, J.; Molnos, S.; Laimighofer, M.; Adamski, J.; Suhre, K.; Strauch, K.; Peters, A.; Gieger, C.; Langenberg, C.; Stewart, I. D.; Theis, F. J.; Grallert, H.; Kastenmüller, G.; Krumsiek, J. Characterization of Missing Values in Untargeted MS-Based Metabolomics Data and Evaluation of Missing Data Handling Strategies. *Metabolomics* **2018**, *14* (10), 128.
(32) 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* (1), 663.
(33) *Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies | Journal of Proteome Research*. https://pubs.acs.org/doi/full/10.1021/acs.jproteome.5b00981 (accessed 2024−10−25).
(34) Lee, K. J.; Carlin, J. B.; Simpson, J. A.; Moreno-Betancur, M. Assumptions and Analysis Planning in Studies with Missing Data in Multiple Variables: Moving beyond the MCAR/MAR/MNAR Classification. *Int. J. Epidemiol.* **2023**, *52* (4), 1268−1275.
(35) *Compositional Data Analysis | Annual Reviews*. https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-042720-124436 (accessed 2024−10−25).
(36) Misra, B. B. Data Normalization Strategies in Metabolomics: Current Challenges, Approaches, and Tools. *Eur. J. Mass Spectrom.* **2020**, *26* (3), 165−174.
(37) Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Orešič, M. Normalization Method for Metabolomics Data Using Optimal Selection of Multiple Internal Standards. *BMC Bioinformatics* **2007**, *8* (1), 93.
(38) Aron, A. T.; Petras, D.; Schmid, R.; Gauglitz, J. M.; Büttel, I.; Antelo, L.; Zhi, H.; Nuccio, S.-P.; Saak, C. C.; Malarney, K. P.; Thines, E.; Dutton, R. J.; Aluwihare, L. I.; Raffatellu, M.; Dorrestein, P. C. Native Mass Spectrometry-Based Metabolomics Identifies Metal-Binding Compounds. *Nat. Chem.* **2022**, *14* (1), 100−109.
(39) Fonville, J. M.; Carter, C.; Cloarec, O.; Nicholson, J. K.; Lindon, J. C.; Bunch, J.; Holmes, E. Robust Data Processing and Normalization Strategy for MALDI Mass Spectrometric Imaging. *Anal. Chem.* **2012**, *84* (3), 1310−1319.
(40) Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B Methodol.* **1982**, *44* (2), 139−160.
(41) Kaul, A.; Mandal, S.; Davidov, O.; Peddada, S. D. Analysis of Microbiome Data in the Presence of Excess Zeros. *Front. Microbiol.* **2017**, *8*. DOI: 10.3389/fmicb.2017.02114.
(42) Quinn, T. P.; Erb, I.; Gloor, G.; Notredame, C.; Richardson, M. F.; Crowley, T. M. A Field Guide for the Compositional Analysis of Any-Omics Data. *GigaScience* **2019**, *8* (9), giz107.
(43) Gloor, G. B.; Macklaim, J. M.; Pawlowsky-Glahn, V.; Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **2017**, *8*, 2224.
(44) Lin, W.; Ji, J.; Su, K.-J.; Qiu, C.; Tian, Q.; Zhao, L.-J.; Luo, Z.; Wu, C.; Shen, H.; Deng, H. omicsMIC: A Comprehensive Benchmarking Platform for Robust Comparison of Imputation Methods in Mass Spectrometry-Based Omics Data. *NAR Genomics Bioinforma.* **2024**, *6* (2), lqae071.
(45) Martino, C.; Morton, J. T.; Marotz, C. A.; Thompson, L. R.; Tripathi, A.; Knight, R.; Zengler, K. A Novel Sparse Compositional Technique Reveals Microbial Perturbations. *mSystems* **2019**, *4* (1), No. e00016.
(46) Gauglitz, J. M.; West, K. A.; Bittremieux, W.; Williams, C. L.; Weldon, K. C.; Panitchpakdi, M.; Di Ottavio, F.; Aceves, C. M.; Brown, E.; Sikora, N. C.; Jarmusch, A. K.; Martino, C.; Tripathi, A.; Meehan, M. J.; Dorrestein, K.; Shaffer, J. P.; Coras, R.; Vargas, F.; Goldasich, L. D.; Schwartz, T.; Bryant, M.; Humphrey, G.; Johnson, A. J.; Spengler, K.; Belda-Ferre, P.; Diaz, E.; McDonald, D.; Zhu, Q.; Elijah, E. O.; Wang, M.; Marotz, C.; Sprecher, K. E.; Vargas-Robles, D.; Withrow, D.; Ackermann, G.; Herrera, L.; Bradford, B. J.; Marques, L. M. M.; Amaral, J. G.; Silva, R. M.; Veras, F. P.; Cunha, T. M.; Oliveira, R. D. R.; Louzada-Junior, P.; Mills, R. H.; Piotrowski, P. K.; Servetas, S. L.; Da Silva, S. M.; Jones, C. M.; Lin, N. J.; Lippa, K. A.; Jackson, S. A.; Daouk, R. K.; Galasko, D.; Dulai, P. S.; Kalashnikova, T. I.; Wittenberg, C.; Terkeltaub, R.; Doty, M. M.; Kim, J. H.; Rhee, K. E.; Beauchamp-Walters, J.; Wright, K. P.; Dominguez-Bello, M. G.; Manary, M.; Oliveira, M. F.; Boland, B. S.; Lopes, N. P.; Guma, M.; Swafford, A. D.; Dutton, R. J.; Knight, R.; Dorrestein, P. C. Enhancing Untargeted Metabolomics Using Metadata-Based Source Annotation. *Nat. Biotechnol.* **2022**, *40* (12), 1774−1779.
(47) Di Guida, R.; Engel, J.; Allwood, J. W.; Weber, R. J. M.; Jones, M. R.; Sommer, U.; Viant, M. R.; Dunn, W. B. Non-Targeted UHPLC-MS Metabolomic Data Processing Methods: A Comparative Investigation of Normalisation, Missing Value Imputation, Transformation and Scaling. *Metabolomics* **2016**, *12* (5), 93.
(48) Hughes, R. A.; Heron, J.; Sterne, J. A. C.; Tilling, K. Accounting for Missing Data in Statistical Analyses: Multiple Imputation Is Not Always the Answer. *Int. J. Epidemiol.* **2019**, *48* (4), 1294−1304.
(49) Melnik, A. V.; da Silva, R. R.; Hyde, E. R.; Aksenov, A. A.; Vargas, F.; Bouslimani, A.; Protsyuk, I.; Jarmusch, A. K.; Tripathi, A.; Alexandrov, T.; Knight, R.; Dorrestein, P. C. Coupling Targeted and Untargeted Mass Spectrometry for Metabolome-Microbiome-Wide Association Studies of Human Fecal Samples. *Anal. Chem.* **2017**, *89* (14), 7549−7559.
(50) Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T. S.; Bushuiev, R.; Murray, K. J.; Hoffmann, N.; Lu, M.; Sarvepalli, A.; Zhang, Z.; Fleischauer, M.; Dührkop, K.; Wesner, M.; Hoogstra, S. J.; Rudt, E.; Mokshyna, O.; Brungs, C.; Ponomarov, K.; Mutabdžija, L.; Damiani, T.; Pudney, C. J.; Earll, M.; Helmer, P. O.; Fallon, T. R.; Schulze, T.; Rivas-Ubach, A.; Bilbao, A.; Richter, H.; Nothias, L.-F.; Wang, M.; Orešič, M.; Weng, J.-K.; Böcker, S.; Jeibmann, A.; Hayen, H.; Karst, U.; Dorrestein, P. C.; Petras, D.; Du, X.; Pluskal, T. Integrative Analysis of Multimodal Mass Spectrometry Data in MZmine 3. *Nat. Biotechnol.* **2023**, *41* (4), 447−449.
(51) Adusumilli, R.; Mallick, P. Data Conversion with ProteoWizard msConvert. *Methods Mol. Biol. Clifton NJ.* **2017**, *1550*, 339−368.
(52) Dixon, P. VEGAN, a Package of R Functions for Community Ecology. *J. Veg. Sci.* **2003**, *14* (6), 927−930.
(53) Kowarik, A.; Templ, M. Imputation with the R Package VIM. *J. Stat. Softw.* **2016**, *74*, 1−16.
(54) *MissForest—non-parametric missing value imputation for mixed-type data | Bioinformatics | Oxford Academic*. https://academic.oup.com/bioinformatics/article/28/1/112/219101 (accessed 2024−10−25).

https://doi.org/10.1021/jasms.4c00434
J. Am. Soc. Mass Spectrom. XXXX, XXX, XXX−XXX