

Article

rox: A Statistical Model for Regression with Missing Values

Mustafa Buyukozkan, Elisa Benedetti and Jan Krumsiek * 

Institute for Computational Biomedicine, Department of Physiology and Biophysics Weill Cornell Medicine, New York, NY 10021, USA

* Correspondence: jak2043@med.cornell.edu

Abstract: High-dimensional omics datasets frequently contain missing data points, which typically occur due to concentrations below the limit of detection (LOD) of the profiling platform. The presence of such missing values significantly limits downstream statistical analysis and result interpretation. Two common techniques to deal with this issue include the removal of samples with missing values and imputation approaches that substitute the missing measurements with reasonable estimates. Both approaches, however, suffer from various shortcomings and pitfalls. In this paper, we present “rox”, a novel statistical model for the analysis of omics data with missing values without the need for imputation. The model directly incorporates missing values as “low” concentrations into the calculation. We show the superiority of rox over common approaches on simulated data and on six metabolomics datasets. Fully leveraging the information contained in LOD-based missing values, rox provides a powerful tool for the statistical analysis of omics data.

Keywords: missing values; regression analysis; limit-of-detection

1. Introduction

High-dimensional molecular datasets, such as metabolomics, proteomics, glycomics, and microbiomics, typically contain a substantial amount of “missing values”, that is, measurement points for which the experimental platform did not return any quantified value [1–3]. Any analysis workflow applied to data with missing values needs to deal with this issue, since most common statistical approaches do not allow for the absence of data points. Missing values in omics data usually occur due to abundances below the instrument sensitivity, the so-called limit of detection (LOD) [3] (Figure 1A). In addition to the obvious loss of information, the presence of missing values interferes with distributional assumptions for statistical analysis. For example, metabolomics measurements are generally log-normally distributed [4], and therefore LOD-based missing values will obfuscate the lower tail of the distribution. In microbiome data, which are compositional in nature [5], left truncation will lead to an artificial overrepresentation of the most abundant species. Further complicating the issue, we have previously shown that LOD effects are not always strict and can occur in blurry fashion, where lower concentration values increase the chance of a value being reported as missing rather than depending on a strict threshold [3].

A statistical method for the analysis of molecular data should take into consideration the abovementioned issues and properties of missing values. First, it should make use of the fact that a missing value indicates a “low” abundance value, even if the precise numeric value is unknown. This allows fully leveraging the information available in the dataset. Ideally, the method should also work in the presence of a non-strict LOD mechanism. Second, in order to be applicable to a wide variety of molecular data, the method should be free of distributional assumptions and robust to outliers.

Existing statistical methods dealing with LOD-based missing values do not or only partially fulfill these requirements. The most popular approaches fall into one of three categories. (1) Missing values are simply deleted from the dataset, which is commonly referred to as complete case analysis (CCA) [6]. Since all samples with any missing values



Citation: Buyukozkan, M.; Benedetti, E.; Krumsiek, J. *rox*: A Statistical Model for Regression with Missing Values. *Metabolites* **2023**, *13*, 127.

<https://doi.org/10.3390/metabo13010127>

Received: 26 October 2022

Revised: 15 November 2022

Accepted: 17 November 2022

Published: 13 January 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/>).

are removed, CCA often leads to a severe reduction in the statistical power, especially when multivariate statistical methods are used. Moreover, if there is an enrichment of missing values in one of the analyzed groups, e.g., in sick individuals compared to healthy ones, CCA will substantially distort the statistical analysis and produce erroneous results [3]. (2) Imputation approaches reconstruct a full data matrix by replacing missing values with reasonable substitutes. “Minimum imputation” is a widely used approach that replaces missing values with the lowest observed value in the data, half of that value, or with a known LOD value [3]. Notably, this approach uses the information that missing values are low but leads to a substantial distortion of the distribution of the analyte [7]. Other common approaches, such as k-nearest-neighbor (knn) imputation, use the correlation structure of the data to infer the original value [8]. These approaches do not use the LOD information and require a strong correlation structure among variables to work properly. (3) The third approach is to use statistical methods that directly incorporate the knowledge of the LOD effect, where missing values are treated as a “low” category. The approach published by [9] addresses the problem of LOD-based left censoring in measurement data using methods from survival analysis, which we prove later in this paper is equivalent to using rank statistics on minimum imputed data. Other approaches make specific assumptions about the underlying data distribution (e.g., log-normal [10] or gamma [11]), and treat the missing values as left-truncated data points from that respective distribution. While incorporating the LOD information, these methods also require strong assumptions on the overall data distribution, which might not be appropriate for certain data types.

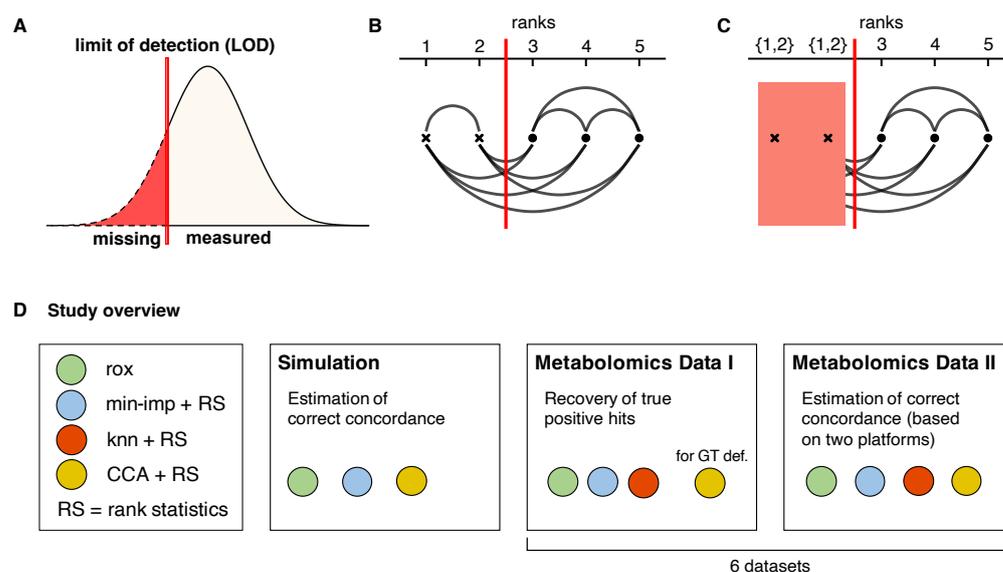


Figure 1. Limit of detection (LOD)-based missingness, statistical concept, and study overview. (A): Schematic of a strict LOD effect on the distribution of a measurement. Values below the LOD (red line) will be reported as missing. (B): Relative order of data points based on their true value. The red line indicates the theoretical LOD. (C): Observed ordering of data points after LOD censoring. While observations below the LOD (red line) cannot be compared once they are censored, we still retain the information that all points below the LOD are lower than all points above the LOD. (D): Overview of *rox* benchmarking. We assess the performance of our approach using an extensive simulation framework, followed by two test scenarios of ground-truth recovery on a series of metabolomics datasets. “for GT def.” = used for definition of ground truth.

Here we present *rox*, “rank order with missing values(X)”, a flexible, nonparametric approach for regression analysis of a dependent variable with missing values and continuous, ordinal, or binary explanatory variables. The core idea is to utilize the knowledge of missing values representing low concentrations due an LOD effect, without requiring any actual imputation steps. The approach is based on rank statistics related to Somer’s D and

Kendall's tau [12–14], and can be computed even with partially quantitative measurements (Figure 1B,C). Leveraging the properties of rank statistics, this framework is applicable to data from any distribution and is robust to outliers. Moreover, while the method relies on the assumption of an LOD effect in its core, it flexibly generalizes to data with other missingness mechanisms.

In this paper, we showcase the features of *rox* on simulated data and benchmark its performance on six real molecular datasets. We use metabolomics data, which is known to be heavily affected by LOD-based missingness and therefore constitutes an optimal test case for this approach. Notably, both for the simulated data and the real data, we define a ground truth for unbiased evaluation. Our analysis demonstrates the superiority of our approach over three of the most commonly used approaches in the field, namely complete case analysis (CCA), minimum imputation and knn-based imputation, coupled with rank-based statistical testing (Figure 1D). Our *rox* implementation is available as open source R package at <https://github.com/krumsieklab/rox>.

2. Methods

2.1. *rox* Core Model

rox is inspired by the ranking-based, nonparametric correlation measure *concordance index* or *c-index* [15], which is equivalent to an ROC-AUC with a binary outcome [16]. Let $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be a set of n observations of two random variables X and Y . A pair of observations $\langle i, j \rangle = \{(x_i, y_i), (x_j, y_j)\}$ is said to be *concordant* if the pairwise ranking of (x_i, x_j) and (y_i, y_j) is the same, i.e., if $\text{sgn}(x_i - x_j) = \text{sgn}(y_i - y_j)$; otherwise, it is said to be *discordant*. The *c-index* is then defined as the number of concordant pairs over the number of all pairs $c_{YX} = \frac{\text{concordant}}{\text{concordant} + \text{discordant}} = \frac{\text{concordant}}{\text{all pairs}}$ [15]. Note that in c_{YX} , the ties in Y are dropped from the calculation, whereas in X , they are counted as 0.5, i.e., neither concordant nor discordant.

The *rox* statistic is an extension of this *c-index* concept to left-censored data. This type of data occurs, for example, when all values below a certain threshold (the *Limit Of Detection*, or LOD) are returned as missing. Based on this LOD assumption, we know that any missing value is lower than any measured value in the data. Importantly, a missing and a measured value can thus still be ranked and are hence *comparable*, while two missing values have no known order and are *noncomparable*. The *rox* method assesses the fraction of concordant pairs only in relation to the *comparable pairs*, $\frac{\text{concordant}}{\text{comparable pairs}}$, a concept that is also used in survival analysis [15]. In the following, we formulate a concordance-based test on comparable pairs.

Let Y now be a left-censored random variable with LOD-based missing values, and let X be an outcome of interest to be associated with Y . As outlined above, any pair of observations $\langle i, j \rangle$ where at least one of y_i or y_j is non-missing constitutes a comparable pair, since we know that the LOD requires, by design, any non-missing value to be larger than all missing values (see also, Figure 1B,C). Let $\pi = \{\langle i, j \rangle \mid y_i \neq \text{NA} \text{ or } y_j \neq \text{NA}\}$ be the set of these comparable pairs, where NA represents a missing value, and let $\Gamma(\pi)$ be the number of concordant pairs in π given X and Y . The nonparametric *rox* coefficient of concordance between variables X and Y , with Y subject to LOD-based missingness, can be formulated as follows:

$$rox_{core} = \frac{\Gamma(\pi)}{|\pi|} \text{ with } \Gamma(\pi) = \sum_{\langle i, j \rangle \in \pi} 1_{\text{if } \text{sgn}(x_i - x_j) = \text{sgn}(y_i - y_j)} \quad (1)$$

where $|\pi|$ is the set size of π , i.e., the total number of comparable pairs.

In general, the *sgn* operator is not defined for missing values. However, under strict-LOD assumptions, any missing value in Y will be lower than all non-missing values. This means that $\text{sgn}(y_i - y_j)$ is always defined in this framework, even when either y_i or y_j is missing. Note that, similar to the *c-index*, the *rox* statistic represents the probability of

concordance between X and Y , and it can take values in the interval $[0,1]$, where 0.5 indicates random ordering, 1 represents perfect concordance, and 0 represents perfect discordance.

2.2. Debiased Weighted Rox Model

In Equation (1), pairs where both y_i and y_j are missing constitute noncomparable pairs, which are excluded from the concordance estimation. However, ignoring these pairs leads to an overall overestimation of positive concordance (>0.5) and an underestimation of negative concordance (<0.5); see remark at the bottom of Supplementary Text S2. Note that the 0.5 cut point is due to the scale of concordance between 0 and 1, where values above 0.5 indicate positive correlation and values below 0.5 represent negative correlation.

To address this problem, we propose a strategy to debias the *rox* coefficient by down-weighting the contribution of missing observations to the overall concordance. To this end, we split all comparable pairs from Equation (1) into two distinct sets: π_b (bridge pairs), which includes pairs of observations where either y_i or y_j is missing, and π_1 , which includes pairs where both y_i and y_j are non-missing (see Supplementary Figure S6). This way, all comparable pairs are partitioned as $\pi = \pi_b \cup \pi_1$.

With this formulation, we can now introduce a weight parameter p to control the contribution of the pairs with missing values, π_b , to the overall *rox* statistics as:

$$rox = \frac{\Gamma(\pi_1) + p\Gamma(\pi_b)}{|\pi_1| + p|\pi_b|}, \quad (2)$$

where $0 \leq p \leq 1$ (see Supplementary Text S1 for a detailed derivation). Setting $p = 1$ leads to the original formulation from Equation (1), which is based on a strict LOD assumption, whereas $p = 0$ reduces the test statistics to a nonparametric complete case analysis, ignoring the contribution of all pairs including any missing values.

In general, if n_0 is the number of missing values, n_1 is the number of non-missing values, and $n = n_0 + n_1$ is the total number of observations, a higher fraction of missing values n_0/n will introduce more bias into the concordance estimation and hence require a lower value of the weight p to debias the estimate. In Supplementary Text S3, we demonstrate that the concordance can be effectively debiased by using the weight factor $p = n_1/n$. With this new expression for p , Equation (2) becomes:

$$rox = \frac{\Gamma(\pi_1) + \frac{n_1}{n}\Gamma(\pi_b)}{|\pi_1| + \frac{n_1}{n}|\pi_b|} \quad (3)$$

2.3. Self-Adjusting Rox for Partial LOD and Non-LOD

The weighted formulation in Equation (3) assumes that the missingness in Y occurs due to a strict LOD threshold, i.e., that *all* values below the LOD threshold will be missing and *all* values above the threshold will be present. However, in many real data scenarios, missingness patterns occur on a continuum [3], from a strict LOD mechanism, to a more probabilistic setting, where lower values have a higher likelihood of being missing, all the way to missing-at-random (MAR). For the *rox* statistic, non LOD-based missingness constitutes a source of bias that will affect the estimation of the true concordance. For cases where the missingness pattern is only marginally due to LOD or even LOD-independent, ignoring the missing values and switching to a complete case analysis is more appropriate.

We thus formulated a self-adjusting version of *rox*. First, we estimated whether the missingness pattern in the data was consistent with an LOD assumption. Let $d_1 = \Gamma(\pi_1)/|\pi_1|$ and $d_b = \Gamma(\pi_b)/|\pi_b|$ be the concordances of pairs with no missing values and pairs with one missing value, respectively. Under strict LOD, which corresponds to a left truncation of the data distribution, if the true concordance is larger than 0.5, then it holds that $d_1 < d_b$, while for concordance values less than 0.5 it holds that $d_b < d_1$ (see Supplementary Text S2 for proof). For simplicity, we only describe the positive concordance case here; the negative concordance case can be derived analogously.

For any random variable Y , we can assess whether the LOD assumption is violated by checking whether $d_1 < d_b$. If the inequality holds, rox concordance is estimated using Equation (3); if it does not, p in Equation (2) is set to zero, removing all missing observations from the analysis and effectively computing the concordance based only on the observations with no missing values, reducing the approach to a complete-case-analysis (CCA).

Thus, the final formulation of rox for positive concordance is:

$$rox = \frac{\Gamma(\pi_1) + \Gamma(\pi_b)p_{d_1 < d_b}}{|\pi_1| + |\pi_b|p_{d_1 < d_b}} \quad (4)$$

where $p_{d_1 < d_b} = p = n_1/n$, as in Equation (3), if $d_1 < d_b$; otherwise, it is 0.

2.4. rox -Based Semiparametric Multivariable Model

The rox model handles one-to-one relations between two variables. In this section, we extend the approach to a semiparametric multivariable modeling framework, which allows modeling the relations between one variable with missing values and multiple variables.

The proposed extension is obtained via the multivariable modeling of the concordance probabilities with an exponential link function [17]. Let Y be a metabolite measurement with missing values and $\mathbf{X} = \{X_1, X_2, \dots, X_k\}$ be k different variables of interest. First, we define the likelihood of concordance for a single pair of observations $\langle i, j \rangle$ as $L_{\langle i, j \rangle} = P(y_i < y_j \mid \langle i, j \rangle) = e^{\eta_i} / (e^{\eta_i} + e^{\eta_j})$, where $\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$ is the standard linear predictor function for sample i , and $\beta = \{\beta_1, \beta_2, \dots, \beta_k\}$ is the vector of the corresponding regression coefficients. The log-likelihood $\ell(\pi)$ of the associated joint probability of all realized pairwise rankings in π can then be formulated as the product of all individual likelihoods:

$$\ell(\pi) = \log(L(\pi)) = \log \left(\prod_{\forall \langle i, j \rangle \in \pi} L_{\langle i, j \rangle} \right) = \sum_{\forall \langle i, j \rangle \in \pi} (\eta_i - \log(e^{\eta_i} + e^{\eta_j})). \quad (5)$$

Similar to the univariate case, Equation (5) ignores pairs of observations where both Y values are missing. However, ignoring these observations leads to a biased estimate of the concordance probability. In this case, we can also debias the model by downweighting the contribution of the missing values and accounting for non-LOD scenarios. We again partition the observation pairs in π into those between two observations with no missing value π_1 and those where one of the two observations is missing π_b (i.e. $\pi = \pi_1 \cup \pi_b$), we introduce a weight p to control the contribution of pairs with missing values, and we check for the violation of the LOD assumption based on the $d_1 < d_b$ inequality:

$$\ell_{rox}(\pi = \pi_1 \cup \pi_b) = \ell(\pi_1) + p_{d_1 < d_b} \ell(\pi_b), \quad (6)$$

where $p = n_1/n$, as derived in Supplementary Text S3. All β coefficients are fitted using a maximum likelihood estimation (MLE) approach, based on a FORTRAN implementation for concordance regression adapted from [17]. The overall concordance of the model is then calculated by computing the rox statistic from Equation (4) between Y and the fitted score $\hat{\eta} = \mathbf{X}\hat{\beta}$.

2.5. Hypothesis Testing

In the following, we derive a hypothesis test for the univariate version of rox , assessing whether $H_0 : rox(Y, X) = 0.5$ can be rejected. Under the null hypothesis of the independence of Y and X , the distribution of the quantity $2 \times rox - 1$ has an expected value of zero. A significance test for rox can be obtained via: $z = (rox - 0.5) / \sqrt{Var(rox)}$, and the corresponding p -value can be calculated via a z-test. The variance of the concordance can be estimated in two different ways: (1) using an equivalent time-dependent Cox model ($cvar$) [16] or (2) through an unbiased infinitesimal jackknife variance estimator ($ivar$) [18].

As pointed out by Therneau et al. [18], *cvar* is an unbiased estimator for $d = 0.5$, while it overestimates the variance if $d \neq 0.5$. On the other hand, *ivar* is unbiased for $d \neq 0.5$, but it underestimates the variance if d is close to 0.5. Taking these findings into consideration, we calculate the p -value of the estimated *rox* statistics based on the average of these two variance estimates, namely, $z = (rox - 0.5) / \sqrt{(cvar + ivar) / 2}$. This approach was inspired by [19], where overestimated and underestimated variances were averaged to obtain a better estimate.

In a multivariable setting, hypothesis testing for the overall model is performed as described in the previous paragraph, with $H_0 : rox(Y, X\beta) = 0.5$, where $\beta = \{\beta_1, \beta_2, \dots, \beta_k\}$ are the regression coefficients, and $X\beta$ is the linear predictor of the model. Furthermore, to test the significance of individual variables in the model, we can use the coefficients in the proposed semiparametric model. In this case, the null hypothesis is defined based on the coefficients: $H_0 : \beta_i = 0$ for variable i . To assess the significance, we used the implementation of [17] to estimate the coefficients and standard errors to calculate a Wald's test [20].

2.6. Simulation Framework

Two continuous variables Y and X with predefined concordance values were simulated. However, concordance cannot be directly parameterized and needs to be determined empirically. Here, we generated the desired concordance by tuning an association parameter as follows: The variable X and a noise term ϵ were first sampled from a standard normal distribution. Y was then defined as $Y = X + \lambda \cdot \epsilon$, where λ determines the association between Y and X . Larger values of λ lead to lower concordance between the two variables. We ranged λ from 0 to 0.7 in steps of 0.01 until the desired concordance $d(Y, X)$ was reached.

For the large sample size simulation, we generated a total of $n = 10,000$ samples. For the small sample size simulation, we first generated a large dataset of $n = 1,000,000$ samples, from which 100 random samples were drawn 1000 times.

In the multivariable case, we simulated a variable Y , an outcome of interest X , and a covariate Z . Correlations between X , Y , and Z were simulated as follows: X was sampled from a normal distribution. A correlated covariate Z was simulated as $Z = X + \epsilon_z$, with ϵ_z being a normal distribution. Y was sampled as $Y = X + Z + 3\epsilon_y$, with ϵ_y again being a normally distributed error term.

2.7. Metabolomics Datasets

To illustrate the performance of *rox* on real data, we analyzed a total of seven previously published metabolomics datasets (Table 1). For the QMDiab plasma validation cohort (HD4), only samples and metabolites overlapping with the HD2 platform were considered. Except for the HAPO dataset, for which only preprocessed data were available, all datasets were preprocessed using the R package *maplet* [21] as follows: Prior to the statistical analysis, the raw peak intensities were normalized using the probabilistic quotient approach [22], using only metabolites with less than 20% missing values to generate the reference sample. Normalized metabolite values were subsequently \log_2 transformed. The following imputation step was applied for all datasets. For each cohort, two additional data matrices were generated: one where missing values were imputed using the minimum value per metabolite and one where missing values were imputed using *knn*-based imputation with 10 neighbors and variable preselection based on pairwise correlation (threshold of 0.2), according to [3].

Table 1. Overview of the metabolomics datasets.

| Cohort | Number of Samples (Controls/Cases) | Number of Metabolites | Phenotype | Specimen | Reference |
|--------------------------------|------------------------------------|-----------------------|-----------------|---------------|-----------|
| QMDiab-Plasma (HD2) | 358 (177/181) | 758 | Type 2 Diabetes | Blood | [23] |
| QMDiab-Urine | 360 (174/186) | 891 | Type 2 Diabetes | Urine | [23] |
| QMDiab-Saliva | 330 (171/159) | 602 | Type 2 Diabetes | Saliva | [23] |
| BRCA | 132 (65/67) | 536 | Breast Cancer | Breast Tissue | [24] |
| RCC | 276 (138/138) | 877 | Kidney Cancer | Kidney Tissue | [25] |
| HAPO | 115 (67/48) | 49 | Hyperglycemia | Plasma | [26] |
| QMDiab-Plasma Validation (HD4) | 292 (137/155) | 359 | Type 2 Diabetes | Plasma | [27,28] |

3. Results

3.1. Simulation Results: Strict LOD

The *rox* method uses rank-based statistics to model measurements with limit of detection (LOD)-based missing value patterns, utilizing the information that absent data points represent low concentrations. It models the relationship between a measurement with missing values as the dependent variable and one or more continuous, ordinal, or binary explaining variables with no missing values. The approach furthermore implements a self-adjusting feature, which detects cases of non-LOD missingness, in which it switches to complete case analysis (CCA). A detailed mathematical derivation of the approach and its properties was provided in the Section 2.

To show how *rox* performed at correctly recovering the true concordance, we developed an extensive simulation framework with a known ground truth. The performance of *rox* was compared to that of complete case analysis (CCA) and regular concordance calculation after minimum imputation (min-imp). Note that the k-nearest-neighbor (knn) imputation was omitted for this part, since it is only feasible in a multivariate setup, where simulation is dependent on various design choices and could easily be tweaked for a method to outperform the others. knn imputation is evaluated based on real datasets later. In the first simulation, a single variable Y and a continuous outcome X were simulated with the concordance d ranging between 0.55 and 0.85. These predefined concordance values represented the ground truth used to evaluate the performance. The missing values were introduced into Y using a strict LOD mechanism, i.e., by setting all values below a given threshold as missing (Figure 2A). We simulated a variety of scenarios by ranging the proportion of missing values in Y from 0% to 90%. For each combination of the true concordance and missing value proportion, we computed the concordance between Y and X by: (i) computing the *rox* statistic, (ii) imputing missing values in Y with minimum value imputation and performing regular concordance analysis, and (iii) only considering complete cases without missing values and performing regular concordance analysis. A large sample size of $n = 10,000$ was chosen to ensure stable results. All simulations were repeated for smaller sample sizes, which yielded equivalent results (see Supplementary Figure S1).

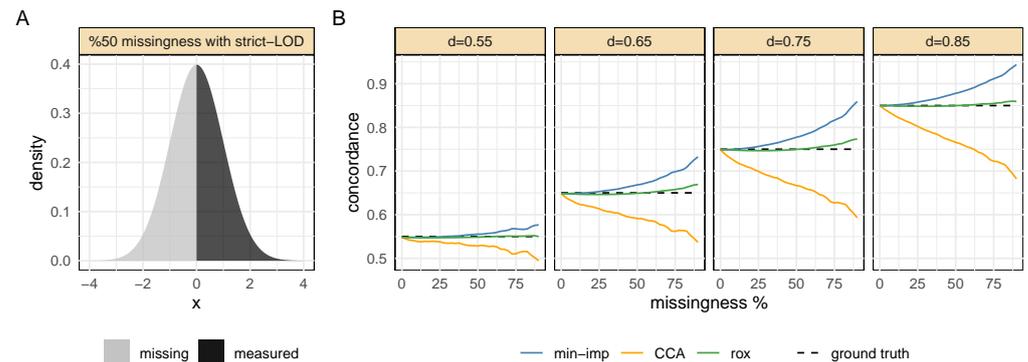


Figure 2. Simulation with the strict LOD mechanism. (A): Example distribution of a simulated variable with 50% missingness due to a strict LOD effect. (B): *rox* outperformed the CCA and minimum imputation in recovering the true concordance across various ground truth values d and missingness fractions. The minimum imputation led to an overestimation of the concordance between the variable and outcome, while the CCA resulted in an underestimation.

The results of this first simulation demonstrated that *rox* was consistently better at retrieving the true concordance than its two competitors (Figure 2B). The minimum imputation generally led to an overestimation of the concordance, while CCA led to an underestimation. With increasing proportions of the missing values, these deviations increased substantially, while the *rox* estimates remained stable and accurate. The same effect was observed across all values of the true concordance values d .

3.2. Simulation Results: Probabilistic LOD

In the second simulation scenario, we evaluated the performance of *rox* in the case of a more realistic “probabilistic LOD” [3], where instead of a hard LOD threshold, the probability of a value being missing continuously increases with decreasing true abundance. A probabilistic LOD was simulated using a sigmoid probability density function that modeled the likelihood of a value being missing given its true value (Figure 3A). The shape of the sigmoid function was parametrized with a variable $pLOD$, which controlled the type of missingness pattern in the data. $pLOD = 0$ led to missing at random (MAR) [29], while $pLOD = 1$ generated a strict LOD effect. Therefore, higher values of $pLOD$ led to more prominent censoring effects (Figure 3B). In this scenario, we again simulated two random variables Y and X with true concordance values ranging from $d = 0.55$ to $d = 0.85$ and $pLOD$ values ranging from 0 to 1. For this setup, the proportion of missing values in Y was fixed at 50%.

The *rox* method again outperformed the competitor methods in all scenarios (Figure 3C) due to the adaptive nature of the model. For simulations with $pLOD$ values below 0.7, *rox* determined that no sufficiently strong LOD effect was present and thus switched to complete case analysis (CCA). Minimum imputation, on the other hand, consistently underestimated the concordance in the range of d between 0 and 0.7, due to its strict implicit LOD assumption. For $pLOD > 0.7$, *rox* leveraged the left-censoring effect and consistently produced more accurate results than its competitors. In this high $pLOD$ range, minimum imputation overestimated the concordance, and the performance of the complete case analysis progressively deteriorated. This behavior was further exacerbated at increasing values of the true concordance d .

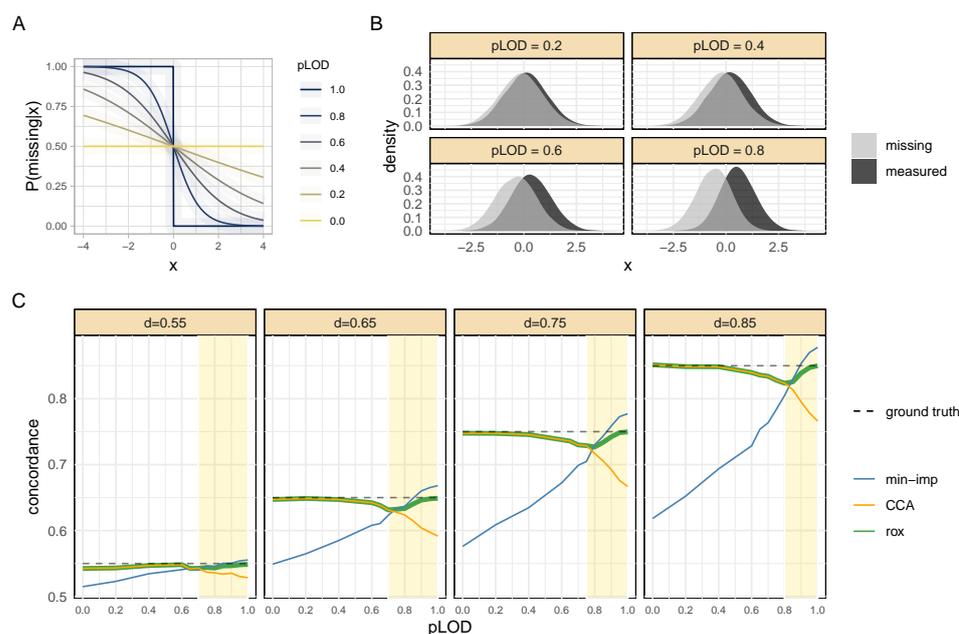


Figure 3. Simulation with the probabilistic LOD mechanism. (A): The probability function describing the likelihood of a value being missing as a function of its numerical value. A $pLOD$ value of 1 results in a strict LOD effect, while $pLOD = 0$ results in values missing at random (MAR), and all values in between produce probabilistic LOD effects. (B): Illustration of the different missingness patterns induced by varying $pLOD$ values. (C): The data were simulated with different $pLOD$ and concordance values, while the missingness percentage was kept at 50% in all scenarios. The yellow shaded area marks the region where our adaptive method automatically identified an active LOD effect and switched from CCA to *rox* analysis. Overall, *rox* outperformed minimum imputation and complete case analysis in all simulation settings.

3.3. Simulation Results: Multivariable Setting

In a third simulation, we investigated how the *rox* model performed in a multivariable setting for both strict-LOD and probabilistic-LOD scenarios. The multivariate setting is of particular interest when the inclusion of multiple variables in the same model is required, for example for covariate correction purposes. In this scenario, we simulated a continuous outcome, a covariate, and a variable of interest under various LOD and missingness settings and again compared *rox*'s performance with that of minimum imputation and knn imputation followed by regular concordance analysis (see Supplementary Figure S2). The results and conclusions were equivalent to the results of univariate simulations suggesting that the *rox* recovered the ground truth better than the competing approaches.

3.4. Evaluation on Metabolomics Data: Recovering High-Confidence Hits

After evaluating the performance of *rox* in a simulation setting, we tested the approach in a real data scenario using published metabolomics datasets from a series of case-control studies (Table 1). In this case, we sought to determine how many true associations between metabolites and the respective study outcomes (e.g., disease status) *rox* could identify compared to its competitor methods.

Defining a ground truth in real datasets, however, is inherently difficult, since a list of true associations is usually not available. For our evaluation framework, we thus constructed a set of metabolite-outcome associations with high confidence of being actual true positives (HC-hits). This set was defined by combining the significant associations obtained from two statistical approaches that are well-suited to detect associations in data with antithetical missing values mechanisms. (1) CCA with the Wilcoxon rank-sum test, which compares two sample groups by filtering out all samples with missing values. This

test works well if values are missing at random and thus do not originate from an LOD effect; however, it is generally underpowered since it entirely excludes missing values from the analysis. (2) Fisher's exact test, which assesses the proportion of missing values in one sample group versus the other, ignoring the actual numeric measurement values. In an LOD setting, this test works well for cases of extreme sample separation, for example, when all samples in one of the comparison groups are low and fall below the LOD threshold. Notably, both approaches suffer from a substantial number of false negatives, since neither is ideally fit for the analysis of molecular data with missing values; however, both methods have very low false positive rates, meaning that the hits they identify are very likely to be correct.

In the following, we used the fraction of HC-hits that each method was able to retrieve as an evaluation metric. The analysis was performed on six datasets: plasma, urine, saliva metabolomics from the QMDiab study [23], where the outcome was type-2-diabetes (T2D), a hyperglycemia study in pregnant women (HAPO) [26], with fasting plasma glucose (low FPG vs. high FPG) as the outcome, and two tissue metabolomics datasets, one from breast tissue (BRCA) [24] and one from kidney tissue (RCC) [25], where the outcome was the origin of the sample (tumor or adjacent-normal tissue) of the sample. Across all datasets, the *rox* outperformed or tied with the other methods in recovering the HC-hits at various significance levels (Figure 4).

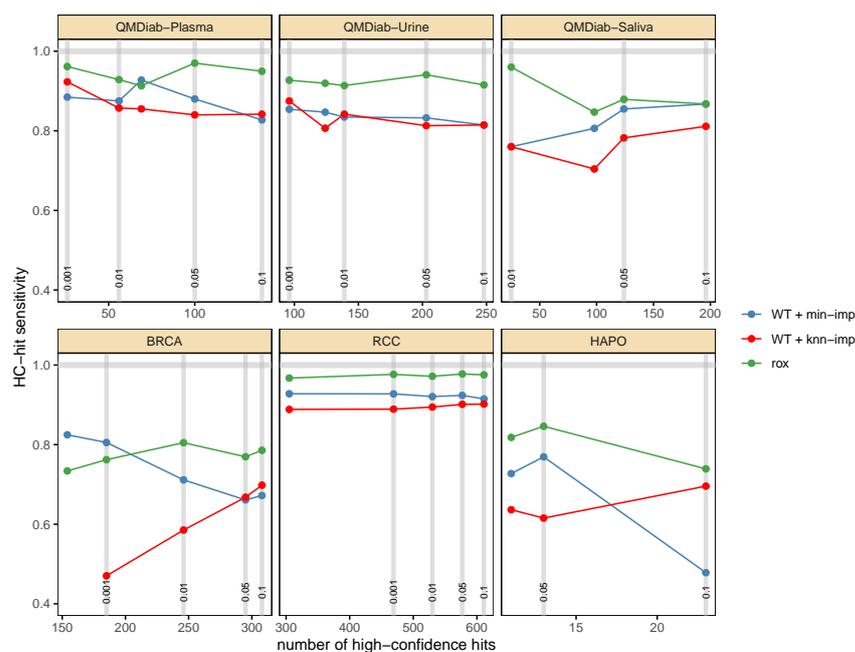


Figure 4. Recovery of high-confidence (HC) hits in six metabolomics datasets. The x-axis shows the number of HC-hits identified with the corresponding Bonferroni-adjusted p -value cutoff. WT: Wilcoxon rank-sum test. The y-axis represents the percentage of HC-hits that were identified, which is equivalent to a measure of sensitivity, at the respective cutoff. The *rox* outperformed or tied with the two imputation approaches across all datasets and cutoffs.

3.5. Evaluation on Metabolomics Data: Multiplatform Validation

A second line of validation on real data was performed on a dataset from the QMDiab study, where the same metabolites were measured in the same samples using two different metabolomics platforms. Specifically, we analyzed metabolites that were fully quantified (FQ) on one platform but were partially missing (PM) on the other platform. We used the concordance between the study outcome and the FQ metabolites without missing values as our ground truth and evaluated the performance of each method based on the consistency between this ground truth and the concordance estimate between the PM metabolite with missing values and the hits at the same outcome. In an ideal scenario, these two

concordance values would be the same, indicating that the method recovered the correct value even in the presence of missing values. To analyze a sufficient number of FQ–PM metabolite pairs, we allowed up to 5% missing values in the FQ candidate and deleted those missing values in the subsequent analysis. If both platforms showed less than 5% missingness for a metabolite, we picked the one with the lower number of missing values as the FQ metabolite and the respective other measurement as the PM metabolite.

Notably, the missing values in these platforms were mostly due to a prominent LOD effect, which we confirmed by comparing the missing and quantified values within the same metabolites across the two platforms (see Supplementary Figure S3). Thus, we expected this dataset to provide a favorable setting for minimum imputation, which assumes a strict LOD. *knn* imputation, on the other hand, cannot impute values outside of the observed data distribution and is therefore unlikely to perform well in a strong LOD scenario [30].

Association analyses were performed between both the FQ and PM metabolites and the respective QMdiab study outcomes (age, sex, BMI, and diabetes), using the *rox* test as well as regular association analysis with the Wilcoxon rank-sum test after *min-imp*, *knn-imp*, and *CCA*. The PM-based concordance values were then compared with the ground truth concordance obtained from the corresponding FQ metabolite. An example of the results for age and PM metabolites with 20% or more missingness is shown in Figure 5A. This analysis was systematically repeated for varying fractions of missingness in the PM metabolite (see Figure 5B and Supplementary Figure S4 for more detailed results). For all outcomes, *rox* was substantially more consistent than minimum imputation, regardless of missingness percentage of the PM metabolite. Notably, the performance of minimum imputation worsened with increasing missingness, while *rox*'s performance remained stable. The *rox* outperformed *knn*-imputation specifically in the association with age, sex, and BMI, while the two methods were mostly comparable for diabetes. Taken together, *rox* performed equivalent to or better than *knn*-imputation across all scenarios. Similar comparisons were performed using multivariable *rox*, with analogous results (see Supplementary Figure S5).

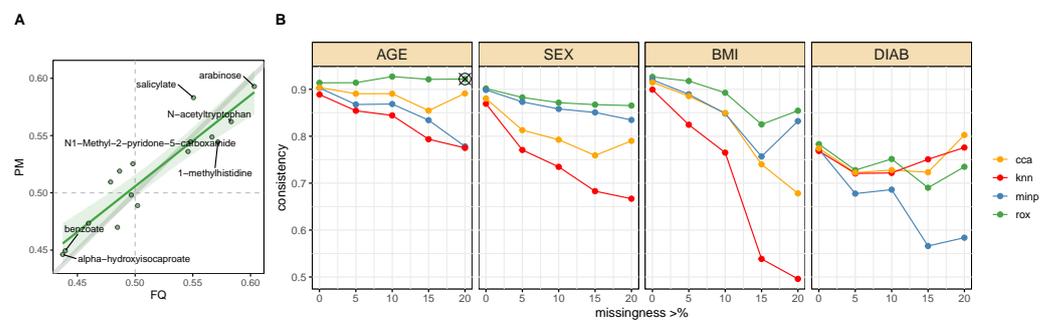


Figure 5. Validation on the two-platform comparison of fully quantified (FQ) and partially missing (PM) metabolites. Plasma metabolites from the same samples were measured on two different platforms. Consistency of the concordance estimates between the FQ-based ground truth and the PM-based estimates was computed for all considered outcomes (age, BMI, diabetes (DIAB), and sex). (A): Consistency across platforms for age associations calculated for PM metabolites with 20% or more missingness. Points distributed along the diagonal indicate consistent estimations between the two platforms. The gray line indicates the $x = y$ axis, while the green line indicates a linear fit over the points of comparison. (B): Systematic results for all four outcomes and varying levels of missingness in the PM metabolites. The x-axis indicates the minimum fraction of missing values per metabolite. The y-axis shows the Pearson correlation between the estimates across the two platforms. Results shown in panel A correspond to the point marked with the black cross. Across all missingness percentages, *rox* was more consistent compared to *knn* imputation and comparable to or better than minimum imputation.

4. Discussion

This paper introduced *rox*, a novel statistical framework for datasets with missing values occurring due to a limit-of-detection (LOD) effect. In contrast to the more common approach of imputing missing values, which comes with various data analysis-related issues, *rox* directly utilized the information that missing values had “low” concentrations. The nonparametric model was based on pairwise ranks and was thus robust to outliers. The method allowed for multivariable modeling and can be used for any quantitative or semiquantitative measurements. Importantly, while *rox* was inherently designed for data with an LOD effect, it also worked with less strict blurry LOD-based data or even when the values were missing at random, in which case it automatically switched to complete case analysis. Using a simulation framework as well as metabolomics datasets from various sample types with different outcomes, we systematically demonstrated the superiority of our method over other approaches that are commonly used in the field. Specifically, *rox* showed higher accuracy in reconstructing the underlying true concordance values and displayed higher statistical power retrieving associations with study outcomes. Notably, while most other studies on real data artificially introduce missing values to evaluate the performance of their statistical approach (e.g., [8,31]), we relied on two data-driven frameworks to define a ground truth for a more realistic evaluation.

In conclusion, we recommend using *rox* for any dataset where an LOD effect can be suspected, even if the effect is not strict. The LOD assumption commonly applies to metabolomics data, as shown in this paper, but it has also been described in data with similar dropout mechanisms, such as proteomics data [32], glycomics data [33], and microbiomics data [34].

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/metabo13010127/s1>, Text S1: Derivation of weighted formulation; Text S2: Proof - LOD leads to $d_b > d > d_1$; Text S3: Choice of weight parameter p for debiasing; Figure S1: Simulation framework and small sample size; Figure S2: Simulation in multivariable setting; Figure S3: Estimating degree of LOD effect in the QMDiab data; Figure S4: Multi-Platform Validation - detailed results; Figure S5: Multi-Platform Validation-multivariable model; Figure S6: Comparable pairs with left-censoring

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this study is freely available. The git repository contains the corresponding links <https://github.com/krumsieklab/rox>.

Conflicts of Interest: J.K. holds equity in Chymia LLC and IP in PsyProtix and is cofounder and equity holder of iollo, Inc.

References

1. Jin, L.; Bi, Y.; Hu, C.; Qu, J.; Shen, S.; Wang, X.; Tian, Y. A comparative study of evaluating missing value imputation methods in label-free proteomics. *Sci. Rep.* **2021**, *11*, 1–11. [[CrossRef](#)] [[PubMed](#)]
2. Lin, H.; Peddada, S.D. Analysis of microbial compositions: A review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes* **2020**, *6*, 1–13. [[CrossRef](#)]
3. Do, K.T.; Wahl, S.; Raffler, J.; Molnos, S.; Laimighofer, M.; Adamski, J.; Suhre, K.; Strauch, K.; Peters, A.; Gieger, C.; et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. *Metabolomics* **2018**, *14*, 128. [[CrossRef](#)]

4. Suhre, K.; Shin, S.Y.; Petersen, A.K.; Mohny, R.P.; Meredith, D.; Wägele, B.; Altmaier, E.; Deloukas, P.; Erdmann, J.; Grundberg, E.; et al. Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **2011**, *477*, 54–60. [CrossRef] [PubMed]
5. 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. [CrossRef] [PubMed]
6. White, I.R.; Carlin, J.B. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat. Med.* **2010**, *29*, 2920–2931. [CrossRef]
7. Helsel, D.R. Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* **2006**, *65*, 2434–2439. [CrossRef]
8. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525. [CrossRef]
9. Helsel, D.R. *Nondetects and Data Analysis. Statistics for Censored Environmental Data*; Wiley-Interscience: Hoboken, NJ, USA, 2005.
10. Moulton, L.H.; Halsey, N.A. A Mixture Model with Detection Limits for Regression Analyses of Antibody Response to Vaccine. *Biometrics* **1995**, *51*, 1570. [CrossRef]
11. Richardson, D.B. Effects of Exposure Measurement Error When an Exposure Variable Is Constrained by a Lower Limit. *Am. J. Epidemiol.* **2003**, *157*, 355–363. [CrossRef]
12. Kendall, M.G. Rank and Product-Moment Correlation. *Biometrika* **1949**, *36*, 177–193. [CrossRef] [PubMed]
13. Newson, R. Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences. *Stata J.* **2002**, *2*, 45–64. [CrossRef]
14. Somers, R.H. A new asymmetric measure of association for ordinal variables. *Am. Sociol. Rev.* **1962**, *27*, 799–811. [CrossRef]
15. Harrell, F.E.; Califf, R.M.; Pryor, D.B.; Lee, K.L.; Rosati, R.A. Evaluating the yield of medical tests. *JAMA* **1982**, *247*, 2543–2546. [CrossRef]
16. Therneau, T.; Atkinson, E. Concordance. en. Vignette of Survival Package. Available online: <https://cran.r-project.org/web/packages/survival/vignettes/concordance.pdf> (accessed on 1 September 2020).
17. Dunkler, D.; Schemper, M.; Heinze, G. Gene selection in microarray survival studies under possibly non-proportional hazards. *Bioinformatics* **2010**, *26*, 784–790. [CrossRef]
18. Therneau, T.M.; Watson, D.A. *The Concordance Statistic and the Cox Model*; Technical Report; Department of Health Science Research, Mayo Clinic: Rochester, MN, USA, 2017; p. 18.
19. Wager, S.; Hastie, T.; Efron, B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* **2014**, *15*, 1625–1651.
20. Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Am. Math. Soc.* **1943**, *54*, 426–482. [CrossRef]
21. Chetnik, K.; Benedetti, E.; Gomari, D.P.; Schweickart, A.; Batra, R.; Buyukozkan, M.; Wang, Z.; Arnold, M.; Zierer, J.; Suhre, K.; et al. maplet: An extensible R toolbox for modular and reproducible metabolomics pipelines. *Bioinformatics* **2022**, *38*, 1168–1170. [CrossRef]
22. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal. Chem.* **2006**, *78*, 4281–4290. [CrossRef]
23. Do, K.T.; Pietzner, M.; Rasp, D.J.; Friedrich, N.; Nauck, M.; Kocher, T.; Suhre, K.; Mook-Kanamori, D.O.; Kastenmüller, G.; Krumsiek, J. Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations. *NPJ Syst. Biol. Appl.* **2017**, *3*, 1–12. [CrossRef]
24. Terunuma, A.; Putluri, N.; Mishra, P.; Mathé, E.A.; Dorsey, T.H.; Yi, M.; Wallace, T.A.; Issaq, H.J.; Zhou, M.; Killian, J.K.; et al. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *J. Clin. Investig.* **2014**, *124*, 398–412. [CrossRef] [PubMed]
25. Hakimi, A.A.; Reznik, E.; Lee, C.H.; Creighton, C.J.; Brannon, A.R.; Luna, A.; Aksoy, B.A.; Liu, E.M.; Shen, R.; Lee, W.; et al. An integrated metabolic atlas of clear cell renal cell carcinoma. *Cancer Cell* **2016**, *29*, 104–116. [CrossRef] [PubMed]
26. Scholtens, D.M.; Muehlbauer, M.J.; Daya, N.R.; Stevens, R.D.; Dyer, A.R.; Lowe, L.P.; Metzger, B.E.; Newgard, C.B.; Bain, J.R.; Lowe, W.L., Jr.; et al. Metabolomics reveals broad-scale metabolic perturbations in hyperglycemic mothers during pregnancy. *Diabetes Care* **2014**, *37*, 158–166. [CrossRef] [PubMed]
27. Do, K.T.; Rasp, D.J.P.; Kastenmüller, G.; Suhre, K.; Krumsiek, J. MoIdentify: Phenotype-driven module identification in metabolomics networks at different resolutions. *Bioinformatics* **2019**, *35*, 532–534. [CrossRef] [PubMed]
28. Mook-Kanamori, D.O.; Selim, M.M.E.D.; Takiddin, A.H.; Al-Homsi, H.; Al-Mahmoud, K.A.; Al-Obaidli, A.; Zirie, M.A.; Rowe, J.; Yousri, N.A.; Karoly, E.D.; et al. 1, 5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. *J. Clin. Endocrinol. Metab.* **2014**, *99*, E479–E483. [CrossRef]
29. Rubin, D.B. Inference and missing data. *Biometrika* **1976**, *63*, 581–592. [CrossRef]
30. Beretta, L.; Santaniello, A. Nearest neighbor imputation algorithms: A critical evaluation. *BMC Med. Inform. Decis. Mak.* **2016**, *16*, 197–208. [CrossRef]
31. Stekhoven, D.J.; Bühlmann, P. MissForest—Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112–118. [CrossRef]
32. Karpievitch, Y.; Stanley, J.; Taverner, T.; Huang, J.; Adkins, J.N.; Ansong, C.; Heffron, F.; Metz, T.O.; Qian, W.J.; Yoon, H.; et al. A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **2009**, *25*, 2028–2034. [CrossRef]

33. Hart, G.W.; Copeland, R.J. Glycomics hits the big time. *Cell* **2010**, *143*, 672–676. [[CrossRef](#)]
34. Silverman, J.D.; Roche, K.; Mukherjee, S.; David, L.A. Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2789–2798. [[CrossRef](#)] [[PubMed](#)]

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.