metabolites

MDPI

Article

Variable Selection in Untargeted Metabolomics and the Danger of Sparsity

Gerjen H. Tinnevelt [1,*], Udo F.H. Engelke [2], Ron A. Wevers [2], Stefanie Veenhuis [3], Michel A. Willemsen [3], Karlien L.M. Coene [2], Purva Kulkarni [2,†] and Jeroen J. Jansen [1,†]

1 Institute for Molecules and Materials, Radboud University, 6525 AJ Nijmegen, The Netherlands; jj.jansen@science.ru.nl

2 Translational Metabolic Laboratory (TML), Department of Laboratory Medicine, Radboud University Medical Center, 6525 GA Nijmegen, The Netherlands; Udo.Engelke@radboudumc.nl (U.F.H.E.); ron.wevers@radboudumc.nl (R.A.W.); karlien.coene@radboudumc.nl (K.L.M.C.); purva.Kulkarni@radboudumc.nl (P.K.)

3 Department of Pediatric Neurology, Radboud University Medical Centre, 6525 GA Nijmegen, The Netherlands; stefanie.veenhuis@radboudumc.nl (S.V.); michel.willemsen@radboudumc.nl (M.A.W.)

* Correspondence: gtinnevelt@science.ru.nl or chemometrics@science.ru.nl

† These authors contributed equally to this work.

Received: 30 September 2020; Accepted: 10 November 2020; Published: 17 November 2020

Abstract: The goal of metabolomics is to measure as many metabolites as possible in order to capture biomarkers that may indicate disease mechanisms. Variable selection in chemometric methods can be divided into the following two groups: (1) sparse methods that find the minimal set of variables to discriminate between groups and (2) methods that find all variables important for discrimination. Such important variables can be summarized into metabolic pathways using pathway analysis tools like Mummichog. As a test case, we studied the metabolic effects of treatment with nicotinamide riboside, a form of vitamin B3, in a cohort of patients with ataxia–telangiectasia. Vitamin B3 is an important co-factor for many enzymatic reactions in the human body. Thus, the variable selection method was expected to find vitamin B3 metabolites and also other secondary metabolic changes during treatment. However, sparse methods did not select any vitamin B3 metabolites despite the fact that these metabolites showed a large difference when comparing intensity before and during treatment. Univariate analysis or significance multivariate correlation (sMC) in combination with pathway analysis using Mummichog were able to select vitamin B3 metabolites. Moreover, sMC analysis found additional metabolites. Therefore, in our comparative study, sMC displayed the best performance for selection of relevant variables.

Keywords: univariate/multivariate statistics; chemometrics; untargeted metabolomics; variable selection; ataxia–telangiectasia; vitamin B3/nicotinamide riboside treatment

1. Introduction

Metabolites are small molecules that are the substrates, intermediates, or end products of metabolism. Alterations in metabolite concentration may be related to a disease or may be caused by xenobiotics [1]. Untargeted metabolomics aims to measure as many metabolites as possible and is often used as a hypothesis-generating or an exploratory tool. One popular analytical technique for untargeted metabolomics is LC/MS, which can generate thousands of features. However, the majority of these features remains unidentified, i.e., their masses do not match any known metabolite present in metabolite databases [2]. Additionally, alterations of a single enzyme may lead to a cascade of secondary metabolic effects, which are measured in untargeted metabolomics but their impact on disease presentation and development is unclear [3].

One goal of untargeted metabolomics is finding metabolite biomarkers predictive of incidence or outcome of disease [4]. Popular chemometric methods for finding biomarkers include univariate analysis such as *t*-tests and also multivariate analysis such as principal component analysis (PCA), partial least squares (PLS), and orthogonal projections to latent structures discriminant analysis (OPLS-DA) [5]. After calculating these multivariate methods, variables may be selected using variable influence on projection (VIP) [6], selectivity ratio (SR) [7], or significance multivariate correlation (sMC) [8]. These methods render many variables of importance. Additionally, sparse methods exist that select variables when creating the model such as lasso [9], sparse PLS (sPLS) [10] and CARS-PLS [11]. These methods find the minimal set of variables needed for an optimal prediction of group membership. All of these methods restrict the number of metabolites included in the differentiating biomarker, where sparse methods have an inherent restriction for a minimal set of contributing metabolites. This complies with Occam's razor that dictates that models should be simplified as much as necessary. However, this does introduce an apparent contradiction, in which the chemical analysis of metabolomics aims for full metabolite coverage and the data analysis aims for optimal simplification. Within metabolic pathways, considerable correlations may exist between different metabolites, imposed by the number of chemical reactions they are involved in. These three aspects together create a discrepancy in which the data analysis may ignore specific metabolites that are being measured in the chemical analysis, because their levels are so highly correlated with the levels of other metabolites, and that are then included in the model. Ignoring such redundant information will not hamper the predictive power of the models, but will seriously limit the mechanistic information that can be obtained from the valuable metabolomics data.

After finding a discriminative set of variables, the associated features must be annotated to actual metabolites to put the results in biological context regarding the disease studied. Mummichog [12] is an algorithm that annotates features to metabolic pathways based on their *m/z* ratio and retention time. If a metabolic pathway contains a high number of significant features, this pathway is most likely affected. Mummichog thus requires to find as many true significant features as possible. Sparse discrimination methods are less suited for Mummichog, because these sparse methods select few features and ignore redundant or correlated features and thus will probably only select a single feature from a pathway to discriminate between groups. Univariate analysis was used before by the authors of Mummichog and, although useful, univariate analysis does not consider the covariance structure present in pathways. Potential metabolites may be missed that do not have a significant difference by themselves, but may in combination with other metabolites (e.g., metabolites in the same pathway) be different between groups. The variable selection method sMC is able to find correlating variables with the response while ignoring the effect of irrelevant correlating structures, a known problem of VIP [8]. For this reason, we hypothesized that the combination of Mummichog with sMC would enhance the possibility of discovering altered metabolic pathways.

In this paper, we compared the performance of multiple variable selection methods in combination with Mummichog for the identification of vitamin B3 metabolites in a cohort of patients with ataxia–telangiectasia (A–T) orally supplemented with vitamin B3 in the form of nicotinamide riboside (for the rationale and study protocol, see: www.clinicaltrials.gov, identifier: NCT03962114). Vitamin B3 is an important co-factor for many cellular processes, including fatty acid metabolism and energy metabolism [13]. Therefore, supplementing vitamin B3 is expected to alter the concentration of many metabolites alongside increasing the levels of vitamin B3 metabolites. The goal of this untargeted metabolomics study was to see whether the vitamin B3 pathway-related metabolites were indeed increased and which other metabolites and/or pathways were influenced upon treatment. This study is thus an excellent showcase for the discrepancy between data analysis (optimal simplification) and untargeted metabolomics (full metabolite coverage).

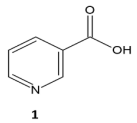
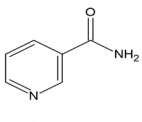
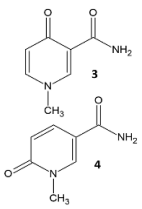
2. Results

The results are subdivided into four sections. Section 2.1. shows that during the treatment the vitamin B3 metabolites were indeed increased with a high fold change, a low *p*-value, and a high sMC F-value. Section 2.2. indicates that all methods could easily discriminate between samples taken before and during treatment. Section 2.3. shows that Mummichog could indeed find the vitamin B3 pathway and additional pathways. Finally, Section 2.4. describes correlation analysis that found relationships between metabolites that were missed by pathway analysis. Metabolites assigned by Mummichog are the best possible guesses based on their *m/z* values, and their name was only given in the main text when confirmed by measurement of a model compound in the Ultra-High Performance Liquid Chromatography-Quadrupole Time-Of-Flight Mass Spectrometry (UHPLC-QTOF-MS) setup [14]. The only exception isN1-methyl-4-pyridone-5-carboxamide and N1-methyl-2-pyridone-5-carboxamide.

2.1. Vitamin B3 Pathway-Related Metabolic Features

Nicotinic acid (**1**) degrades via nicotinamide (**2**) into N1-methyl-4-pyridone-5-carboxamide (**3**) or N1-methyl-2-pyridone-5-carboxamide (**4**), see Table 1. Mummichog assigned two features at a weight tolerance of 1 ppm to nicotinic acid or picolinic acid as isobaric metabolites. Nicotinic acid and picolinic acid were both detected according to the Human Metabolome Database (HMDB) [15]. Three metabolic features were assigned to nicotinamide (**2**) at a weight tolerance of 1 ppm. Metabolites **3** and **4** are isomers, and the five features at 1 ppm were assigned to both metabolites. At a weight tolerance of 5 ppm, two more metabolic features were assigned to **1**, five more to **2**, and 23 metabolic features were assigned to metabolites **3** and **4**. Most of these features had different retention times than the features found at 1 ppm except for the M+K[1+] adduct at a retention time of 2.86 minutes and M+CH3COO[−] adduct at 2.97 minutes for metabolite **3** or **4**. Mummichog probably assigns many false positives at 5 ppm and misses some (false negatives) at 1 ppm. Most features and metabolites in Table 1 had a significantly higher intensity after treatment. Although no definitive level 1 identification [16] of these four metabolites could be given, it is highly likely that they indeed represent vitamin B3-related metabolites, increased upon treatment with nicotinamide riboside, also because the *p*-value is low and the sMC F-value is extremely high. A high sMC F-value means the feature explained much variance of the response to treatment compared to other structures in the data.

Table 1. Assigned metabolites of vitamin B3 metabolism: the degradation of nicotinic acid (**1**) into nicotinamide (**2**) and into N1-methyl-4-pyridone-5-carboxamide (**3**) or N1-methyl-2-pyridone-5-carboxamide (**4**) at a weight tolerance of 1 ppm using Mummichog.

Assigned Metabolite	Adduct	*m/z*	Retention Time	Fold Change	Wilcoxon *p*-Value	sMC F-Value
 **1**	M + H[1+]	124.0329	1.20	7.00	1.2×10^{-4}	302.90
	M + H[1+]	124.0329	1.01	2.02	1.2×10^{-4}	118.44
 **2**	M − H_2O + H[1+]	105.0445	3.52	1.06	0.9	0.0149
	M + H[1+]	123.0552	1.38	2.76	1.2×10^{-4}	193.84
	M + H[1+]	123.0553	0.97	1.94	1.2×10^{-4}	254.34
 **3** **4**	M−NH_3 + H[1+]	136.0394	2.86	9.50	1.2×10^{-4}	361.59
	M − H[−]	151.0511	2.97	6.77	1.2×10^{-4}	359.63
	M + H[1+]	153.0659	2.96	7.75	1.2×10^{-4}	377.00
	M + Na[1+]	175.0479	2.86	6.98	1.2×10^{-4}	539.53
	M + Cl[−]	187.0276	2.97	7.20	1.2×10^{-4}	187.52

2.2. Treatment of A–T Patients with Nicotinamide Riboside

Table 2 shows that all tested methods could clearly distinguish the patient samples before and during treatment with nicotinamide riboside. The univariate Wilcoxon sign test and multivariate methods OPLS-DA with VIP and Weight Randomization Test for partial least squares (WRT-PLS) with sMC render hundreds of deviating features, whereas the sparse methods CARS-PLS-DA and sPLS-DA only result in a handful of altered features. A feature with *m/z* 87.0089 ($C_3H_4O_3$ M − H[−]) was increased more than 47-fold upon treatment. Mummichog assigned this feature to eight metabolites and, of the total fifteen features that CARS-PLS-DA found, only two others were assigned to known metabolites. The same features were also found with sPLS-DA, but additionally, eleven more features could be annotated including the M + Cl[−] of metabolite **3** or **4**. All of the 48 selected features had a high positive (>0.63) or high negative (<-.55) correlation with the features of Table 1. Only eight features had an sMC F-value higher than 187.52 and only two metabolic features had an sMC F-value higher than 539.53. Thus, the sparse methods CARS-PLS-DA and sPLS-DA did not only select the features with the highest discriminatory power, see Figure 1.

Table 2. Overview of results of different methods to distinguish patients with A–T before and during treatment with nicotinamide riboside.

Method	Wilcoxon Sign Test with False Discovery Rate Correction	OPLS-DA VIP	WRT-PLS sMC $\alpha = 0.01$ F > 7.06	WRT-PLS sMC $\alpha = 3 \times 10^{-7}$ F > 32.99	CARS-PLS-DA	sPLS-DA
Accuracy	N/A	100%	100%	100%	100%	100%
Number features	612	842	770	214	15	48
Annotated features	51	41	73	23	3	14

The last row of the table shows the number of metabolites annotated by Mummichog.

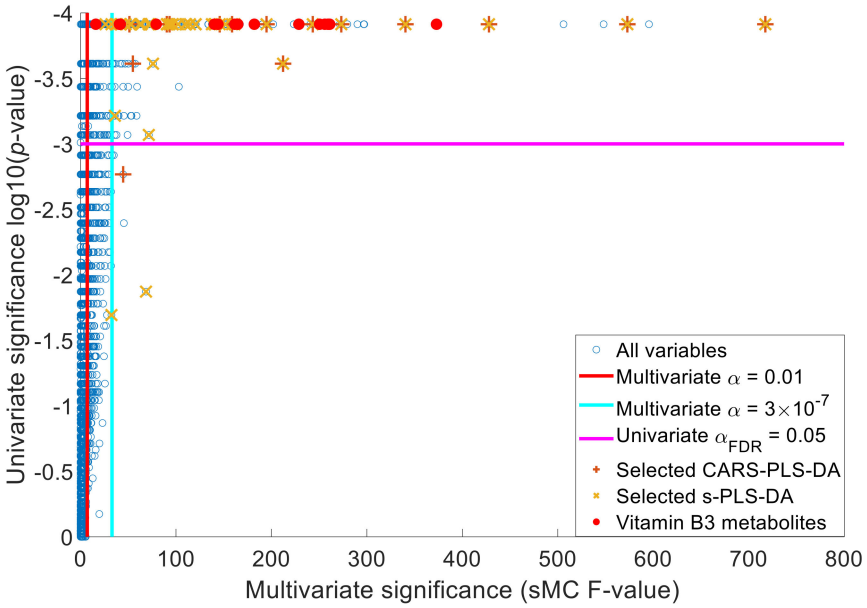


Figure 1. Multivariate significance (sMC F-value) against the univariate significance log10(*p*-value) of the variables (blue unfilled circles). The vitamin B3 metabolites (red filled circles) were both multivariately and univariately significantly different between before and during treatment. Despite selecting many significant variables, the sparse variable selection methods sPLS-DA (yellow crosses (×)) and CARS-PLS-DA (brown plus signs (+)) did not select the vitamin B3 metabolites.

2.3. Pathway Analysis with Mummichog Based on Features Selected with Wilcoxon or sMC

The feature table used for analysis contained a total of 22,684 features. Table 3 shows the number of features assigned to metabolites by Mummichog. Next, Mummichog was used to find significant enriched pathways based on the Wilcoxon sign test and sMC. Pathways may have too many false-positive features assigned to them and permutation testing is used to detect those. The *p*-values were randomly permutated 100 times, each time the significance level was calculated for the pathway, and the part of permutations in which the pathway had a lower significance level than the actual data was calculated. The permutation test results of the vitamin B3 pathway can be seen in Table 3, which shows that only at a weight tolerance level of 1 ppm the pathway is not a false positive, regardless of the approach.

Table 3. Mummichog annotation results and vitamin B3 pathway permutation results.

Tolerance	Features	Metabolites	Ratio	Wilcoxon Sign Test	sMC $\alpha = 0.01$ F > 7.06	sMC $\alpha = 3 \times 10^{-7}$ F > 32.99
1 ppm	1033	618	2.41	0	0	0
3 ppm	2465	959	2.68	0.99	1	0.46
5 ppm	3268	1131	2.68	1	1	0.88

The column Features shows the number of features that could be assigned to a metabolite. The column Metabolites shows the number of unique metabolites detected. The column Ratio shows the average number of metabolites assigned to the same metabolic feature. The columns Wilcoxon sign test, sMC $\alpha = 0.01$, and $\alpha = 3 \times 10^{-7}$ show the permutation value for the vitamin B3 pathway.

The additional pathways identified by Mummichog are listed in Table 4. The analysis based on the *p*-values derived from the Wilcoxon sign test only found the vitamin B3 pathway to be significantly altered, whereas sMC identified additional significant pathways. The metabolic features associated with metabolites of the arachidonic acid pathway were decreased during nicotinamide riboside treatment. Another pathway found with sMC at $\alpha = 0.01$ involved methionine and cysteine metabolism. Adenosine and another metabolite were significantly increased, one metabolite was significantly decreased, and five metabolites showed a significant multivariate correlation with response to the treatment, including cysteine.

Table 4. Pathways found by Mummichog at a weight tolerance of 1 ppm.

Pathway	Wilcoxon Sign Test	sMC $\alpha = 0.01$	sMC $\alpha = 3 \times 10^{-7}$
Vitamin B3	5 *	5 *	4 *
Arachidonic acid	3	6 *	3 *
Methionine and cysteine	4	9 *	2

The table lists the number of significant metabolites found within the pathway. The asterisk (*) means that it is significant according to the more conservative Expression Analysis Systematic Explorer (EASE) test.

2.4. Correlation Analysis with Only the Significant Annotated Metabolites

Finally, pathway analysis does not show any relationship between the vitamin B3 metabolites and other metabolites. For this reason, the correlation matrix was calculated between all significant features annotated with primary ions. An absolute correlation of 0.8 was used as cut-off to find structures in the data. Most features were uncorrelated except for one highly connected network containing 21 features, including the vitamin B3 metabolites. The following assigned metabolites in the network were significantly increased upon treatment with nicotinamide riboside: nicotinamide, nicotinic acid, N1-methyl-4-pyridone-5-carboxamide, N1-methyl-2-pyridone-5-carboxamide, hypoxanthine, guanosine, inosine, adenosine, and two more. Four assigned metabolites were significantly

decreased. The purine metabolism pathway was not significant in the pathway analysis because many non-significant features were also assigned to metabolites from this pathway.

3. Discussion

Untargeted metabolomics with LC/MS identifies thousands of features in patient body fluids. Adequate variable selection is therefore essential to find clinically relevant features. In our study, a cohort of patients with A–T was sampled before and during treatment with nicotinamide riboside. The goal of this study was to compare multiple variable selection methods on their ability to find the nicotinamide riboside-related metabolites and how treatment with vitamin B3, as an important co-factor in human metabolism, would influence other metabolic pathways. All the compared variable selection methods could easily distinguish between before and during treatment patient samples, indicating that the individual variability is much smaller than the difference between before and during treatment. The sparse methods CARS-PLS-DA and sPLS-DA were used to find the minimal set needed for optimal discrimination between before and during treatment. Wilcoxon signed test and partial least squares in combination with VIP and sMC were used to find all discriminatory features. Although no metabolic feature was assigned to nicotinamide riboside, features were assigned to vitamin B3 metabolites. The reason that we do not observe nicotinamide riboside itself as increased may be found in matrix-related ion suppression, as we were able to detect the pure compound when dissolved in water. Another reason could be that nicotinamide riboside itself is metabolized rapidly to the associated metabolites that we do assign.

The vitamin B3-related features were significantly increased and had a low Wilcoxon signed test *p*-value, high fold change, high VIP, and high sMC. However, sparse methods only need a subset of those features and thus may only find features that correlate to the feature of interest. For example, CARS-PLS-DA did not find any metabolic feature related to the vitamin B3 pathway and sPLS-DA only found one metabolic feature. The found features did have a high correlation with the vitamin B3 metabolic features. Moreover, not all found features had the highest multivariate correlation significance. These sparse methods probably use some high discriminatory features to discriminate and some less discriminatory features to stabilize the multivariate space. Variable selection in combination with random forest or support vector machines was briefly explored (results not shown), but we did not find any deterministic results and neither did variable selection based on genetic algorithms, because there are multiple subsets possible of the hundreds of discriminatory features.

Another challenge in untargeted metabolomics is the annotation of features to metabolites. Mummichog can automatically assign features to metabolites that are in known metabolic pathways. At a molecular weight tolerance of 1 ppm, Mummichog only assigned 4.55% of all features. Using a higher molecular weight tolerance leads to more assignments, but also increases the rate of false-positive assignments. Mummichog automatically assigns all features to metabolites based on their m/z value. Mummichog thus considers the ppm value and the number of adducts that were also assigned to the same metabolite [12]. For the metabolites specified in this paper, their identity has been confirmed by measurement of a model compound in the UHPLC-QTOF-MS setup [14]. For all other metabolites, such as the arachidonic acid pathway metabolites, further confirmation of identity is still needed. Our approach should therefore be considered as a hypothesis-generating tool, rendering interesting leads for biochemical follow-up studies.

Pathway analysis of the effects of nicotinamide riboside treatment using sMC as input for the Mummichog algorithm pointed to vitamin B3, arachidonic acid, and methionine and cysteine metabolism to be affected. Metabolites of arachidonic acid were decreased and vitamin B3 is known to inhibit the breakdown of lipids [17]. In another study, methionine and cysteine were found to increase in Sprague Dawley rats upon vitamin B3 intake during three months [18]. However, methionine was not found to be significantly different and cysteine was found only multivariately upon treatment with nicotinamide riboside in patients with A–T. The pathway reference database MetaFishNet is a compilation of multiple databases including the Kyoto Encyclopedia of Genes and Genomes (KEGG)

database [19–21]. Most of the significant features were not included in the methionine and cysteine pathway of the KEGG database. Thus, if just the KEGG database is used for the reference of pathways, methionine and cysteine metabolism would not be significant. In other words, pathway analysis does not consider the importance of certain metabolites inside a pathway. The core metabolites are not necessarily significantly different as seen with the methionine and cysteine pathway. Finally, changing the significance level of sMC to $\alpha = 3 \times 10^{-7}$ found the same number of metabolites as the Wilcoxon signed rank test for the arachidonic acid pathway, but because the total number of significant metabolites was lower, arachidonic acid was a significant pathway when using sMC compared to the Wilcoxon signed rank test. Thus, finding more important features is of less value if these features are not biochemically related and sMC may help to find many correlated features.

4. Materials and Methods

4.1. Measurements and Conversion to the Feature Intensity Table

Plasma samples from 14 patients with A–T before treatment and during treatment with vitamin B3 (nicotinamide riboside) were measured using UHPLC-QTOF-MS in both positive and negative ionization mode according to a previously described procedure [14]. Considering the processing of raw QTOF-MS data, the vendor data format (.d) was converted into open format .mzML data using MSConvert [22]. Data files were preprocessed, which included alignment, peak picking, and grouping using XCMS [23]. All the data conversion and preprocessing steps were performed as a part of an in-house developed bioinformatics pipeline. All patient samples were measured in duplicate, and an internal quality control (QC) sample was also included every 9th or 10th measurement. The use of patient samples for this study was approved by the Regional Committee on Research involving Human Subjects Arnhem-Nijmegen (NL68197.091.18). Informed consent was obtained from all patients and/or their legal caregivers according to the tenets of the Declaration of Helsinki.

4.2. Additional Preprocessing of the Large Feature Intensity Table

The preprocessed output encompasses a large feature intensity table where each feature contains an *m/z* value, a retention time, and an intensity (relative abundance) for every measured sample. Features with more than 20% zeroes (missing values) in the samples or with any zero values in the QC samples were removed. Features with an *m/z* value lower than 70 or higher than 700 were removed. Features with a retention time lower than 0.4 minutes or higher than 16 minutes were removed. Intra-batch correction was performed using the internal QC samples and support vector regression (QC-SVR) [24]. Other preprocessing steps included Probabilistic Quotient Normalization (PQN) using the median of internal QC samples as a reference [25], KNN imputation (k = 10, Euclidian distance) of the zero values (missing values), glog transformation optimized based on the internal QC samples [26], and mean centering (no scaling) [27]. The rationale behind PQN is that the majority of (housekeeping) metabolites do not change. The quotients of most samples were close to 1, meaning that the difference between samples was very small and minimized during the sample preparation. Pareto scaling was explored, but standard PLS was unable to distinguish between before and during treatment patient samples; therefore, no scaling was used. Matlab functions were written and used using Mathworks Matlab R2020a.

4.3. Data Analysis, Variable Selection, and Validation

From each duplicate measurement, the mean was calculated prior to univariate analysis with a Wilcoxon signed rank test. Multiple testing correction was performed using the false discovery rate (FDR) approach of Benjamini–Hochberg [28]. For multivariate analysis, the duplicate measurements were used directly, because multivariate methods may distinguish between variance that explains the response and variance that does not explain the response, e.g., instrumental variability. Multivariate analysis included the following: OPLSDA [29,30] in combination with VIP [6], weight randomization

test for partial least squares (WRT-PLS) [31] in combination with sMC [8], CARS-PLS-DA[11] with two latent variables, and sparse PLS-DA with one latent variable [10]. Double cross-validation was used to validate the results [32]. Five-fold cross-validation with ten iterations was used for the outer loop to determine the accuracy. Four-fold cross-validation was used for the inner loop to determine the number of latent variables in OPLS-DA and the variables in CARS-PLS-DA and sparse PLS-DA. Inner cross-validation was not required in WRT-PLS. Variables were selected based on the training set by VIP, sMC, CARS-PLS-DA, and sparse PLS-DA, and then a new PLS-DA model was made using only those variables and the training set, see Figure 2. Accuracy was calculated based on the prediction of the test set. Features that were selected during all cross-validation iterations were listed per method. Matlab functions were written and used using Mathworks Matlab R2020a.

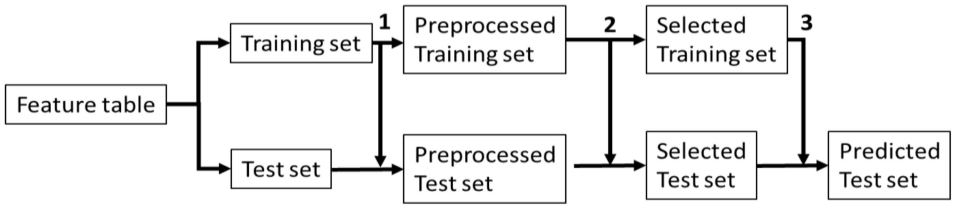


Figure 2. Schematic representation of cross-validation procedure. First, the data were divided into a training set and a test set. (**1**) Preprocessing of data using parameters derived from the training set such as mean and k neighbors. (**2**) Inner cross-validation based on the training set to select variables and number of (orthogonal) latent variables using DQ^2. (**3**) PLS-DA model based on the variable selected training set and used to predict the test set. Based on the predicted test set, an accuracy was calculated.

4.4. Annotation and Pathway Analysis with Mummichog

Mummichog v2.0 was used along with the MS Peaks-to-Pathway module available on Metaboanalyst v4.0 (https://www.metaboanalyst.ca) [12,33]. The input file requires *m/z* values, retention time values, *p*-values, and information on whether the feature is acquired from negative or positive ionization mode. The sMC F values were converted to *p*-values and were used, as well as the *p*-values of the univariate test. The molecular weight tolerance was used at 1, 3, and 5 ppm together with the option to enforce primary ions. Only features with at least one assigned primary ion (M + H[1+], M + Na[1+], M − H_2O + H[1+], M − H[1−], M − 2H[2−], M − H_2O − H[1−]) were valid. The *p*-value cut-off was set to $\alpha = 0.01$ or 3×10^{-7} for sMC and to 0.0012 (FDR-corrected $\alpha = 0.05$) for univariate testing. The human reference pathways of the MetaFishNet database were used for pathway analysis in Mummichog [34].

5. Conclusions

Through untargeted metabolomics, hundreds of altered metabolic features were identified in ataxia–telangiectasia (A–T) patients upon treatment with nicotinamide riboside, which included vitamin B3-associated metabolites. Chemometric methods can easily distinguish between samples obtained before and during treatment; however, assigning all the metabolic features to metabolites and connecting them to disease mechanisms is still a challenging task. Moreover, the definition of a limited set of key metabolites to monitor therapy response is preferred in light of future application in clinical diagnostics.

For this reason, sparse methods were tested that focus on finding a minimum set of metabolic features to discriminate between patient samples before and during treatment. However, these sparse methods did not find vitamin B3 metabolites but only found features that correlate with the vitamin B3 metabolites. Another approach is to summarize the detected metabolites into their related pathways using pathway analysis. Mummichog in combination with *p*-values obtained from univariate testing only found the vitamin B3 pathway. Using Mummichog with significant multivariate correlation based on partial least squares discriminant analysis also found the vitamin B3 pathway and additionally

found the arachidonic acid pathway. The metabolites related to the arachidonic acid pathway still need further confirmation of their identity, to draw definite conclusions. Additional correlation analysis of the significant annotated features found that the purine nucleosides, i.e., adenosine, inosine, and guanosine, correlate to the vitamin B3 metabolites. In conclusion, our data show that treatment with nicotinamide riboside in patients with A–T leads to an increase of vitamin B3 metabolites in plasma as expected, and also increases plasma purine nucleosides and may decrease metabolites in the arachidonic acid pathway. Even though we realize that many metabolic features that may provide additional relevant information still remain unannotated, we show that significant multivariate correlation in combination with pathway analysis and correlation analysis is able to render more leads for putatively affected biochemical processes as compared to sparse methods.

Author Contributions: Conceptualization, G.H.T., M.A.W., K.L.M.C., P.K. and J.J.J.; Data curation, G.H.T., U.F.H.E. and P.K.; Formal analysis, G.H.T.; Investigation, G.H.T. and U.F.H.E.; Methodology, G.H.T., K.L.M.C. and P.K.; Resources, U.F.H.E., S.V. and M.A.W.; Software, G.H.T. and P.K.; Supervision, R.A.W. and J.J.J.; Validation, U.F.H.E.; Visualization, G.H.T.; Writing – original draft, G.H.T.; Writing – review & editing, R.A.W., S.V., M.A.W., K.L.M.C., P.K. and J.J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research made use of metabolomics infrastructure that is part of the NWO-funded Netherlands X-omics initiative, project 184.034.019.

Acknowledgments: We would like to thank Siebolt de Boer and Ed van der Heeft for their technical assistance.

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

References

1. Pezzatti, J.; Boccard, J.; Codesido, S.; Gagnebin, Y.; Joshi, A.; Picard, D.; González-Ruiz, V.; Rudaz, S. Implementation of liquid chromatography–high resolution mass spectrometry methods for untargeted metabolomic analyses of biological samples: A tutorial. *Anal. Chim. Acta* **2020**, *1105*, 28–44. [CrossRef] [PubMed]
2. de Souza, L.P.; Alseekh, S.; Naake, T.; Fernie, A. Mass Spectrometry-Based Untargeted Plant Metabolomics. *Curr. Protoc. Plant Biol.* **2019**, *4*, e20100.
3. Patti, G.J.; Yanes, O.; Siuzdak, G. Metabolomics: The apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–269. [CrossRef] [PubMed]
4. Strimbu, K.; Tavel, J.A. What are biomarkers? *Curr. Opin. Hiv Aids* **2010**, *5*, 463. [CrossRef] [PubMed]
5. Gertsman, I.; Barshop, B.A. Promises and pitfalls of untargeted metabolomics. *J. Inherit. Metab. Dis.* **2018**, *41*, 355–366. [CrossRef]
6. Galindo-Prieto, B.; Eriksson, L.; Trygg, J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS). *J. Chemom.* **2014**, *28*, 623–632. [CrossRef]
7. Rajalahti, T.; Arneberg, R.; Berven, F.S.; Myhr, K.-M.; Ulvik, R.J.; Kvalheim, O.M. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 35–48. [CrossRef]
8. Tran, T.N.; Afanador, N.L.; Buydens, L.M.; Blanchet, L. Interpretation of variable importance in partial least squares with significance multivariate correlation (sMC). *Chemom. Intell. Lab. Syst.* **2014**, *138*, 153–160. [CrossRef]
9. Tibshirani, R. Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2011**, *73*, 273–282. [CrossRef]
10. Lê Cao, K.-A.; Boitard, S.; Besse, P. Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinform.* **2011**, *12*, 253. [CrossRef]
11. Li, H.; Liang, Y.; Xu, Q.; Cao, D. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Anal. Chim. Acta* **2009**, *648*, 77–84. [CrossRef] [PubMed]
12. Li, S.; Park, Y.; Duraisingham, S.; Strobel, F.H.; Khan, N.; Soltow, Q.A.; Jones, D.P.; Pulendran, B. Predicting network activity from high throughput metabolomics. *PLoS Comput. Biol.* **2013**, *9*, e1003123. [CrossRef] [PubMed]
13. Yang, Y.; Sauve, A.A. NAD+ metabolism: Bioenergetics, signaling and manipulation for therapy. *Biochim. Et Biophys. Acta (BBA) Proteins Proteom.* **2016**, *1864*, 1787–1800. [CrossRef] [PubMed]
14. Coene, K.L.; Kluijtmans, L.A.; van der Heeft, E.; Engelke, U.F.; de Boer, S.; Hoegen, B.; Kwast, H.J.; van de Vorst, M.; Huigen, M.C.; Keularts, I.M. Next-generation metabolic screening: Targeted and untargeted metabolomics for the diagnosis of inborn errors of metabolism in individual patients. *J. Inherit. Metab. Dis.* **2018**, *41*, 337–353. [CrossRef]

15. Wishart, D.S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A.C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S. HMDB: The human metabolome database. *Nucleic Acids Res.* **2007**, *35*, D521–D526. [CrossRef]
16. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J.L. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007**, *3*, 211–221. [CrossRef]
17. Zhang, Y.; Schmidt, R.J.; Foxworthy, P.; Emkey, R.; Oler, J.K.; Large, T.H.; Wang, H.; Su, E.W.; Mosior, M.K.; Eacho, P.I. Niacin mediates lipolysis in adipose tissue through its G-protein coupled receptor HM74A. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 729–732. [CrossRef]
18. Basu, T.K.; Makhani, N.; Sedgwick, G. Niacin (nicotinic acid) in non-physiological doses causes hyperhomocysteineaemia in Sprague–Dawley rats. *Br. J. Nutr.* **2002**, *87*, 115–119. [CrossRef]
19. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef]
20. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **2019**, *28*, 1947–1951. [CrossRef]
21. Kanehisa, M.; Sato, Y.; Furumichi, M.; Morishima, K.; Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **2019**, *47*, D590–D595. [CrossRef] [PubMed]
22. Adusumilli, R.; Mallick, P. Data conversion with ProteoWizard msConvert. In *Proteomics*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 339–368.
23. Tautenhahn, R.; Patti, G.J.; Rinehart, D.; Siuzdak, G. XCMS Online: A web-based platform to process untargeted metabolomic data. *Anal. Chem.* **2012**, *84*, 5035–5039. [CrossRef] [PubMed]
24. Kuligowski, J.; Sánchez-Illana, Á.; Sanjuán-Herráez, D.; Vento, M.; Quintás, G. Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC). *Analyst* **2015**, *140*, 7810–7817. [CrossRef] [PubMed]
25. Filzmoser, P.; Walczak, B. What can go wrong at the data normalization step for identification of biomarkers? *J. Chromatogr. A* **2014**, *1362*, 194–205. [CrossRef] [PubMed]
26. Parsons, H.M.; Ludwig, C.; Günther, U.L.; Viant, M.R. Improved classification accuracy in 1-and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinform.* **2007**, *8*, 234. [CrossRef]
27. Di Guida, R.; Engel, J.; Allwood, J.W.; Weber, R.J.; 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*, 93. [CrossRef]
28. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
29. Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119–128. [CrossRef]
30. Bylesjö, M.; Rantalainen, M.; Cloarec, O.; Nicholson, J.K.; Holmes, E.; Trygg, J. OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *J. Chemom.* **2006**, *20*, 341–351. [CrossRef]
31. Tran, T.; Szymańska, E.; Gerretzen, J.; Buydens, L.; Afanador, N.L.; Blanchet, L. Weight randomization test for the selection of the number of components in PLS models. *J. Chemom.* **2017**, *31*, e2887. [CrossRef]
32. Szymańska, E.; Saccenti, E.; Smilde, A.; Westerhuis, J. Double-check: Validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics* **2012**, *8*, 3–16. [CrossRef] [PubMed]
33. Pang, Z.; Chong, J.; Li, S.; Xia, J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. *Metabolites* **2020**, *10*, 186. [CrossRef] [PubMed]
34. Li, S.; Pozhitkov, A.; Ryan, R.A.; Manning, C.S.; Brown-Peterson, N.; Brouwer, M. Constructing a fish metabolic network model. *Genome Biol.* **2010**, *11*, R115. [CrossRef] [PubMed]