Kwaghe International
Journal of Sciences and Technology

ISSN: 1595-7977

Index: Harvard, Boston, Sydney University, Dimensions, Lens, ResearchGet Scilit, Semantic, Google Scholar, Base etc

https://doi.org/10.58578/KIJST.v2i2.6486

Preliminary Study into the Application of *Metabolomics* in Soil Discrimination

Abdulwasiu Olawale Salaudeen[1*], Yemisi Ajoke Olawore[2], Aishat Abdulkareem Yetunde[3], Hajara Yakubu[4], Abubakar Umar Dewa[5]

[1,2,3]National Mathematical Centre Abuja, Nigeria; [4]University of Abuja, Nigeria;
[5]Federal University of Kashere Gombe, Nigeria
abdulwasiu.olawale@gmail.com

Article Info:

Submitted:	Revised:	Accepted:	Published:
May 15, 2025	Jun 13, 2025	Jun 25, 2025	Jun 30, 2025

Abstract

Soil metabolomics provides a comprehensive analysis of small-molecule metabolites (≤1.5 kDa) present in soil and offers insights into how environmental processes influence soil conditions. Although this technique has been applied to various soil-related studies, it remains underrepresented in the broader field of metabolomics, highlighting the need for further research. This study aims to characterize the soil metabolome across contrasting soil sites to evaluate the discriminating capacity of soil metabolomics and its potential as a soil quality indicator. Soil metabolites were extracted using methanol and dichloromethane and analyzed with an Agilent 1260 Infinity II liquid chromatography–mass spectrometry platform. A total of 307 compounds were positively identified, including steroids, saponins, amino acids, organothiophosphorus compounds, and fatty acids. Multivariate statistical tools, such as Partial Least Squares Discriminant Analysis (PLS-DA) score and loading plots, Variable Importance in Projection (VIP) scores, Significance Analysis of Microarrays (SAM), and heat mapping successfully discriminated

the soil samples from four distinct sites. Among the identified metabolites, prolyl-hydroxyproline (ID 1817) had the highest VIP score (≈2.62) and emerged as a potential biomarker for differentiating soil types. These findings underscore the utility of metabolomics in soil characterization and its potential application in environmental monitoring and soil quality assessment.

Keywords: Soil Metabolomics; Environmental Analysis; Metabolites; PLS-DA; Soil Biomarkers

INTRODUCTION

Soil is essential to life as we know it, in addition to supporting a variety of ecosystem services like nutrient cycling, carbon (C) sequestration, climate regulation, and flood mitigation [1–5]. It is fundamental to many ecosystem functions that are necessary for the proper operation of the earth system [6]. However, maintaining soil function over time is a significant problem because of growing agricultural practices, acidification, salinisation, biological invasions, desertification, and urbanisation [7] in addition to weather patterns that are becoming more erratic [8].

The chemical profiling of biologically derived compounds in a variety of species (plants, microbes, algae, etc.) and environmental compartments (soil, water, etc.) is made possible by untargeted metabolomics [9–12]. The goal of this analytical method is to find the greatest number of compounds in the 50–2000 Da range [13-14]. Earthworms, plants, and soil microbial communities have all been subjected to pesticide exposure, and its effects have been studied using metabolomics [3, 10, 15], because it can reveal details regarding the metabolic processes of soil microorganisms [13, 16], along with other omic techniques, it has also been proposed as a potent method to characterise pesticide biodegradation [16].

Recent advancements in spectroscopy have made it possible to identify and measure the relative abundance of hundreds of metabolites found in biological samples [17]. The cost of metabolomics is comparable to that of proteomics and genomics [18]: it permits rapid sample processing [19], and is not constrained by post-translational modifications and varying levels of epigenetic regulation [17]. The method can also be used to find biochemical intermediates in interacting metabolic pathways, which could advance

our knowledge of the biological processes that occur in soil and enhance our capacity to forecast results [20].

The primary force behind soil functioning is now understood to be soil biology. However, it is typically under-represented in soil quality evaluations, possibly due to its tremendous complexity, both in terms of multispecies interaction and interpretation, compared to more standard chemical and physical attributes [6]. Therefore, to completely integrate biological markers into routine soil monitoring, a unified suite of techniques is required. Significant progress has been achieved in assessing the variety of organisms that inhabit soil (for example, through metagenomics and community profiling); yet, it has been challenging to establish a direct correlation between these advancements and variations in the quantity and composition of soil organic matter (SOM) and soil function [4, 21]. There are currently many methods available to characterise the composition of small molecules in soils, such as high-performance liquid chromatography (HPLC) [22], nuclear magnetic resonance (NMR) [3], capillary electrophoresis-mass spectrometry (CE-MS) [23], gas chromatography-mass spectrometry (GC-MS) [14, 24], and Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) [25]. We concentrate on the commonly used GC-MS and liquid chromatography-mass spectrometry (LC-MS) techniques. While GC-MS is a widely used and reasonably priced technique, its detectable metabolite range is somewhat limited, and sample derivatisation is necessary to improve analyte volatility. LC-MS is another widely used technique that is starting to be helpful for soil metabolomics [26]. By comparing the millions of chemical features produced by this method with real chemical standards in terms of retention time, m/z values, and fragmentation spectra, the features can be recognised. In addition, novel compound identification can be achieved by LC-MS, usually in combination with NMR structural studies and isolation [27]. LC-MS and GC-MS techniques can be applied separately, or samples can be fractionated and examined with the help of an additional set of techniques. For this approach, for instance, LC-MS is utilised for the analysis of bigger and more diversified sets of polar and nonpolar metabolites, while GC-MS is employed to identify carbohydrates and minor organic acids.

Soil metabolomics is a developing method for characterising various small molecule metabolites, or metabolomes, in the soil. Fatty acids, amino acids, lipids, organic acids, sugars, and volatile organic compounds are examples of soil metabolites. These chemicals are closely related to soil biogeochemical cycles, which are fuelled by soil microbes, and frequently contain vital elements like sulphur, phosphorus, and nitrogen. Research work in

this area is still limited. In this work we explore the potential of metabolomics in soil chemistry.

MATERIALS AND METHODS

Soil sampling

A set of 24 soils, including 5 blank samples containing no soil, were sampled in four locations on Penang mainland in the morning of April 26, 2024. Over a 7 m altitudinal gradient, six samples were taken from each sample point; the geo-references are provided in Table 1. This region is considered to have a tropical climate. Rice is the principal crop grown at the sample locations. Following the removal of the remaining litter layer, a 25 cm depth sample of the soil was taken, and it was kept in the laboratory at -80°C until analysis.

Table 1: Geo-references of sample sites

Sample location	Geo-reference
A	5.447429, 100.426566
B	5.460760, 100.449149
C	5.464304, 100.449505
D	5.52242, 100.475278

Untargeted metabolomics

The 24 collected soil samples and 5 blank samples containing no soil were lyophilised on an Alpha 1-4 LSC lyophiliser. To minimise metabolite alterations, the lyophilised samples were then kept in separate sterile glass vials at -80 °C [28]. Dichloromethane and methanol were used to extract the soils. For the purpose of metabolomics fingerprinting of soils, methanol is typically utilised as an extraction solvent [3, 13, 29]. Both methanol and dichloromethane were used as extraction solvents because of their contrasting polarities. This was done in an effort to target polar to apolar compounds and broaden the spectrum of extractable metabolites. There were three extractions carried out: two successive extractions with methanol and one with dichloromethane. 0.02 mg/L triphenylmethanol (97%) was used as an internal standard.

For each sample, 0.5 g dry weight of soil was vortexed for 1 minute with 1 mL of methanol and the internal standard, followed by ultrasonication for 45 minutes. The ultrasonicated samples were then centrifuged for 5 minutes at 4000 rpm and 4 °C, and the supernatant was collected. The same sequence was repeated on the residue but with 30

minutes of ultrasonication. Finally, the residue was extracted with 1 mL dichloromethane with 1 minute of vortex, 20 minutes of ultrasonication, and 5 minutes of centrifugation at 4000 rpm and 4 °C. The supernatants from the three extraction cycles were collected, mixed, and filtered with 0.02 μm PTFE filters. Extracts were stored in HPLC amber vials in the freezer at -20 °C until analysis. The same extraction procedure was carried out for the blank without the soil sample.

Using high-resolution mass spectrometry and Agilent 1260 Infinity II liquid chromatography, untargeted metabolomics analysis of soil samples was carried out. Chromatographic separation was performed with a C18 column (100 × 2.1 mm, 1.8 μm) at a flow rate of 0.3 mL/min. The injection volume was 5 μL, and the column was kept at 35 °C. Solvent A (0.1% (v/v) formic acid in water) and solvent B (acetonitrile) were used as the mobile phase, and the gradient was as follows: 0–0.5 min 3% (B), 0.5–1 min 3% (B), 1–9 min 50% (B), 9–13 min 100% (B), 13–14 min 100% (B), 14–14.5 min 3% (B), 14.5–18 min 3% (B). Ionisation was performed in negative mode with a scan range of 50–1500 Da.

Metabolomic data and statistical analysis

Raw.D data were processed using MS-DIAL v 4.90, yielding 11891 RT-m/z features MS DIAL parameters are provided in Table 1.

Table 2: MS-DIAL parameters used

#Project Ionisation type: Soft ionisation Separation type: Chromatography (GC, LC, CE, or SFC) MS method type: Conventional LC/MS or data dependent MS/MS Data type (MS1): Profile Data type (MS/MS): Centroid Ion mode: Negative Target omics: Metablomics					
DATA COLLECTION	**PEAK DETECTION**	**MS2DEC**	**IDENTIFICATION**	**ADDUCT**	**ALIGNMENT**
MS1 tolerance: 0.01	Minimum peak height: 2000	Sigma window value: 0.5	Retention time tolerance: 18	[M-H]⁻ [M-H_2O-H]⁻ [M+Na-2H]⁻ [M+Cl]⁻ [M+K-2H]⁻ [M+FA-H]⁻ [M+Hac-H]⁻ [M+C_2H_3N+Na-2H]⁻ [M+Br]⁻ [M+TFA-H]⁻	Retention time tolerance: 0.05
MS2 tolerance: 0.025	Mass slice width: 0.1	MS/MS abundance cut off: 0	Accurate mass tolerance (MS1): 0.01		MS1 tolerance: 0.015
Retention time begin: 0.1	Smoothing method: LWMA	Exclude after precurso	Accurate mass tolerance (MS2): 0.05		Retention time factor: 0.5

		r ion: √		[M-$C_6H_{10}O_4$-H]- [M-$C_6H_{10}O_5$-H]- [M-$C_6H_8O_6$-H]- [M+CH_3COONa-H]- [2M-H]- [2M+FA-H]- [2M+Hac-H]- [3M-H]- [M-2H]$^{2-}$	
Retention time end: 18	Smoothing level: 3	Keep the isotopic ions until: 0.5	Identification score cut off: 80		MS1 factor: 0.5
MS1 mass range begin: 50	Minimum peak width: 5	Keep the isotopic ions w/o MS2Dec			
MS1 mass range end: 1500			Use retention time for scoring:		Peak count filter: 0%
MS/MS mass range begin: 50			Use retention time for filtering:		N% detected in at least one group: 0%
MS/MS mass range end: 1500			Retention time tolerance: 0.1		Remove features based on blank info.
Maximum charged number: 2			Accurate mass tolerance: 0.01		Sample max/blank average: 5
Consider Cl and Br elements: √			Identification score cut off: 85%		Keep ‘reference matched” metabolite features: √
Number of threads: 4			Relative abundance cut off: 0		Keep ‘suggested (w/o MS2)’ metabolite features:
Execute retention time correction: √			Only report the top hit:		Keep removable features and assign the tag: √
					Gap filling by compulsion: √

LWMA = Linear Weighted Moving Average

Data and statistical analysis

The data was normalised by log10 transformation for all subsequent analysis. Partial Least Squares Discriminant Analysis (PLSDA) was applied as a supervised method of determining variance within and between soil classes. Score plot, loading plot, variable

importance in projection (VIP) plot, violin plot, Significance of microarrays (SAM) and heat map were generated MetaboAnalyst 6.0 platform. Metaboanalyst 6.0 parameter are provided in Table 3.

Table 3: Metaboanalyst 6.0 parameters

Data Filtering	Reliability filter:		RSDs greater than: 25%
	Variance filter	Interquantile range (IQR)	Percentage to filter out: 10%
	Abundance filter	Mean intensity value	Percentage to filter out: 0%
Normalisation Overview	Sample normalisation: Normalisation by median	Data transformation: Log transformation (base 10)	Data scaling: Auto scaling

RESULTS

Figure 1. Partial Least Squares Discriminant Analysis (PLS-DA) overview from metabolomics data of soil samples

This plot shows an overview of the Partial Least Squares Discriminant Analysis (PLS-DA) score derived from the metabolomics data of soil samples. Pairwise component

comparisons are visually displayed, most likely illustrating how several soil samples cluster in the reduced multivariate space.

Components 1 through 5 are the five components displayed. Each element is a linear combination of the initial metabolic variables, or metabolite characteristics, chosen to optimise the degree of difference across soil samples. Despite only accounting for 9.7% of the variance in the dataset, principal component 1 (PC1) explains the most of it, showing a high level of data complexity. PC2, which accounts for 6.6% of the variance, is the second most significant component. The PCs that follow (PC3 to PC5) explain decreasing percentages of the variation, indicating that while they each contribute very little on their own, together they characterise the structure in the data.

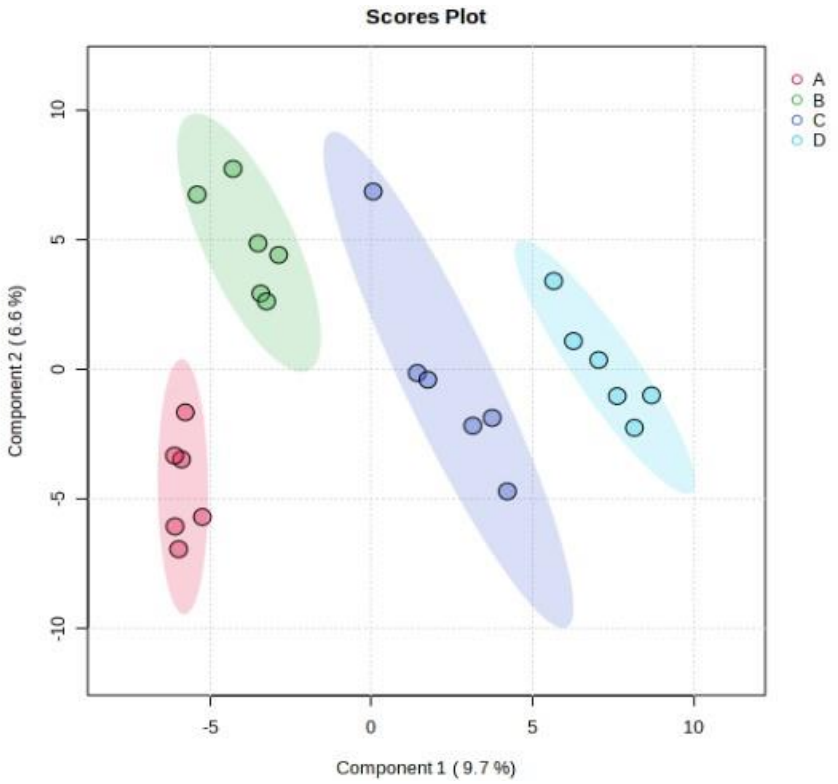

Figure 2. Partial Least Squares Discriminant Analysis (PLS-DA) score plot from soil samples collected at different sites

PLS-DA score plot was employed as a supervised method meant to optimise separation between specified groups of samples based on their metabolomic profiles. Component 1 (shown by the x-axis) accounts for 9.7% of the variation in the metabolomics data that is associated with the sample group separation. Component 2 is represented by the y-axis, which accounts for 6.6% of the variation. When combined, these

two elements account for 16.3% of the data's volatility. PLS-DA concentrates on identifying components that best separate the preset groups (A, B, C, and D) because it is supervised. The colourful ellipses depict confidence intervals (typically 95%) for each group, indicating the distribution and variability within each group.

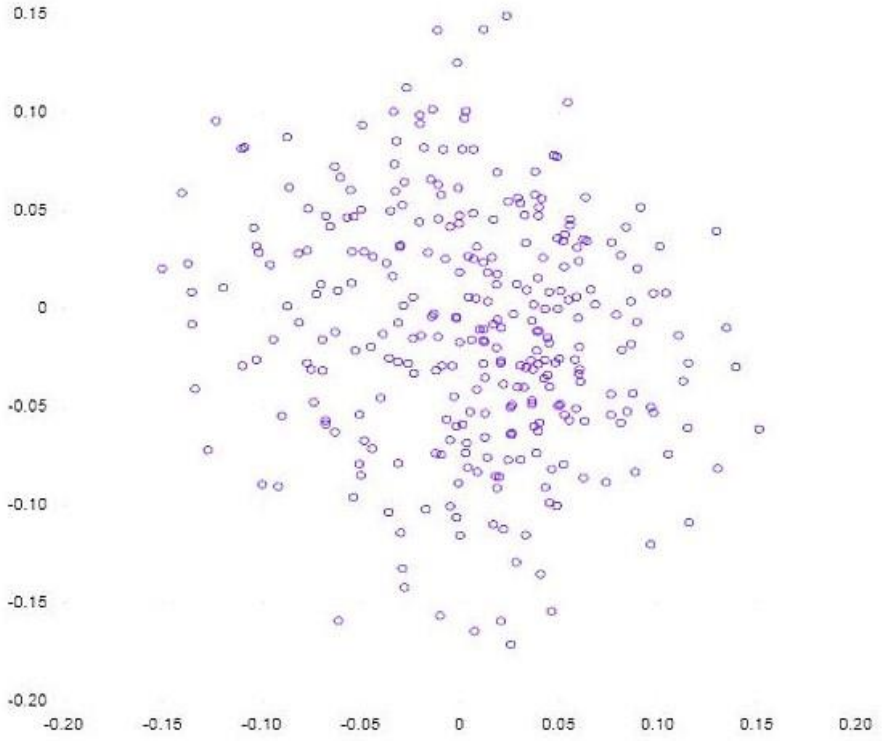

Figure 3. Loading plot for soil metabolomics data

The loading plots show how metabolites affect the analysis's ability to distinguish across sample sites. Every dot on the diagram represents a different metabolite. The dots' placements correspond to the relative contributions of each metabolite to the variation that the model's PCs capture. The model is more affected by metabolites that are further away from the origin, or the centre, and they also contribute more to the variance observed across soil samples taken at various locations. The loadings of the first two main components, or latent variables in PLS-DA, are represented by the X and Y axes. The magnitude of each metabolite's contribution to the dataset's direction of highest variation is shown by these loadings. Since there are both positive and negative loadings in this Figure, it is possible that the trends in the variation of metabolites between groups are at odds.

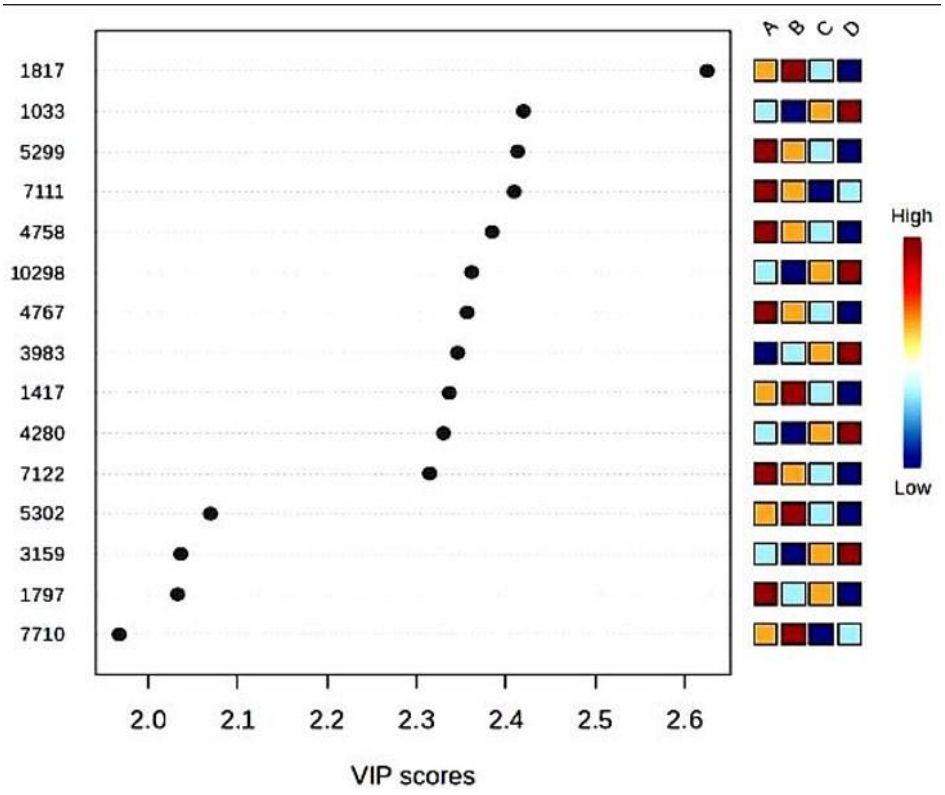

Figure 4. VIP score plot of the metabolomics study of soil samples from different sites

Based on the PLS-DA model, the Variable Importance in Projection (VIP) scores (shown on the X-axis) indicate how important each unique metabolite is in explaining the variability in the dataset. Greater VIP scores signify a metabolite's greater significance in differentiating soil samples from various locations. A VIP score of more than one is generally seen as significant, with higher levels denoting greater relevance. A distinct metabolite, denoted by a numerical ID (e.g., 1817, 1033, 5299, etc.), is represented by each dot on the diagram. The most crucial metabolites for differentiating the soil samples from the various sites are those at the top of the list and with the highest VIP ratings. The plot's right-hand heatmap (A, B, C, D) shows the colour-coded distribution of metabolite abundance in each of the samples. The abundance levels (low to high) of each metabolite are represented by the colour spectrum from blue to red. This colour scheme provides insight into site-specific metabolic variations by making it easier to see which metabolites are more variable across the various sample sites. A distinct sampling location is represented by each column, and each row corresponds to the metabolite on the left.

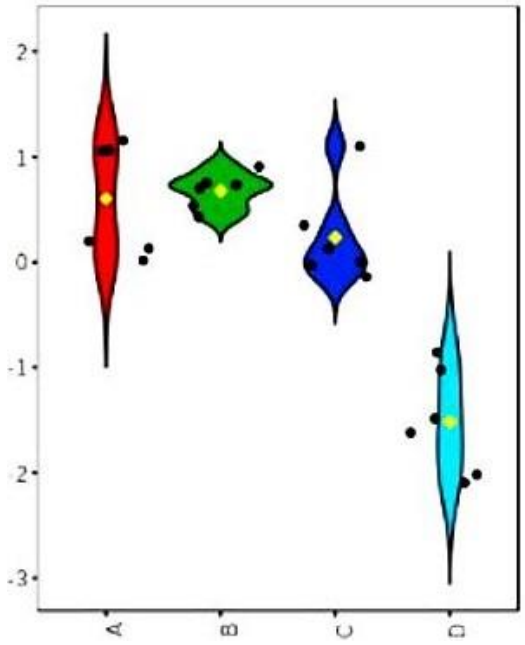

Figure 5. Violin plot of concentration distribution of top VIP (1817), prolyl-hydroxyproline (Pro-Hyp) across four sampling points

Figure 5 shows the concentration distribution of prolyl-hydroxyproline (Pro-Hyp), which is the metabolite (1817) with the top VIP score of ≈2.62 at four sampling points (A, B, C, and D) as a violin plot. Prolyl-hydroxyproline is a dipeptide produced from the amino acids proline and hydroxyproline and is commonly connected with collagen metabolism or soil-derived microbial activities. This metabolite is important in distinguishing between the sampling points because this plot is a component of the Variable Importance in Projection (VIP) score analysis. Following data normalisation, the plot's Y-axis displays the Pro-Hyp concentration values that have been normalised. Positive and negative values represent variations from the average concentration across samples, and the four sampling points (A, B, C, and D) are represented by the X-axis. The distribution and density of Pro-Hyp concentrations within each group are displayed by the violin plot, where larger violin parts denote higher density and more samples with comparable values. The narrow sections highlight areas where the concentrations of fewer samples are similar, suggesting either uncommon or extreme levels. The black points represent individual sample values from each group, and the yellow dots reflect the median Pro-Hyp value for each group.

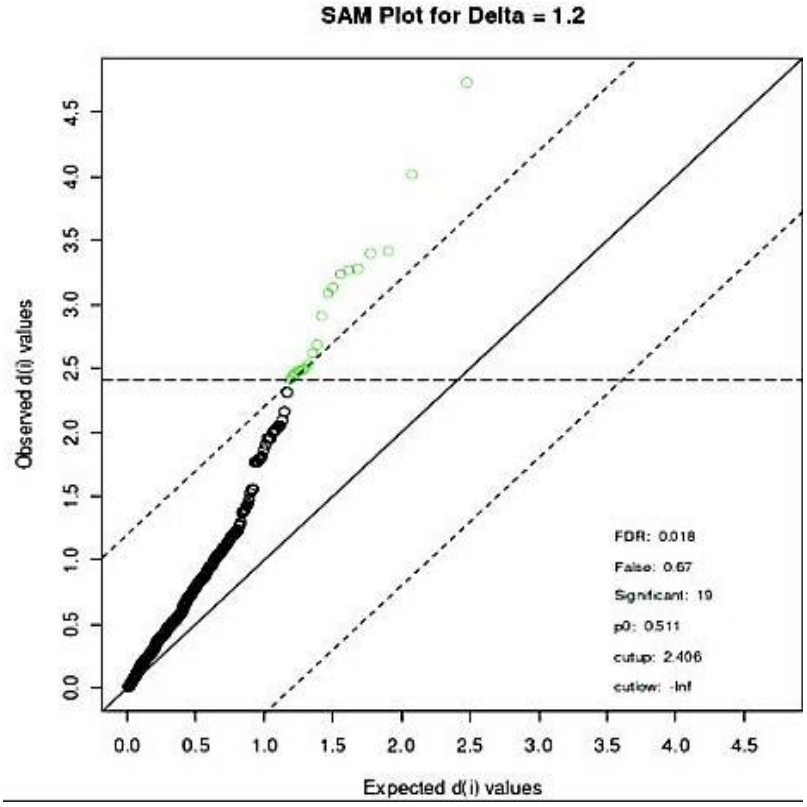

Figure 6. Significance Analysis of Microarrays (SAM) plot for a metabolomics dataset of soil samples

A Significance Analysis of Microarrays (SAM) plot created with MetaboAnalyst 6.0 for a metabolomics dataset is displayed in Figure 6. The Δ value of 1.2, which establishes the threshold for classifying metabolites as significantly different, is the foundation of the Figure. The trade-off between detecting true positives and false positives is managed by the delta threshold. The number of important metabolites that have been identified decreases with a greater delta and increases with a lower delta. The expected d(i) values, or values based on the null hypothesis, are shown on the X-axis. The actual differences seen in the metabolomics dataset are represented on the Y-axis as observed d(i) values. A metabolite is represented by each data point. Points outside and above the dashed lines indicate substantial differences between groups; points below the solid diagonal line indicate no meaningful difference at all. According to the SAM analysis, metabolites shown with green circles are regarded as significantly different and fall outside the dashed diagonal lines.

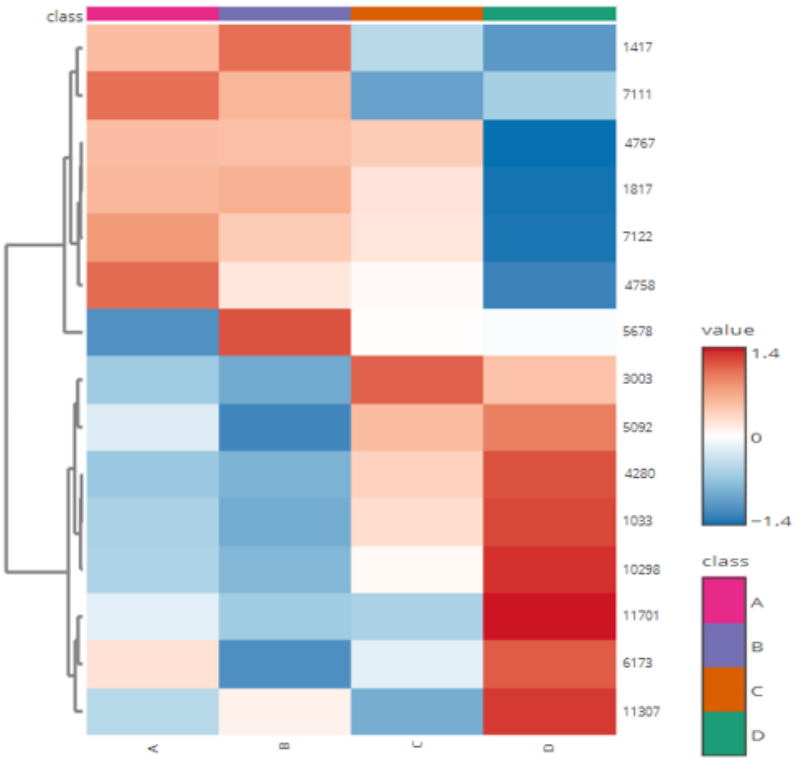

Figure 7. Heat map for the metabolomic data from soil samples collected at different sites.

Heat maps are helpful for displaying the various metabolites' intensity levels across sample sites. This makes it possible to spot trends and group samples according to metabolite abundance. The relative intensity of the metabolite levels is shown by the colour scale on the right. Higher intensity or abundance (positive z-score) is represented by red. Lower intensity or abundance is represented by blue (negative z-scores). Intermediate levels are denoted by white (z-score near to 0). A metabolite is represented by each row, while a distinct soil sample location (A, B, C, or D) is represented by each column. Coloured bars at the top represent the sample locations. The varied colours and patterns of the sample groups point to variations in the metabolomic profiles of the various soil samples. Based on their patterns throughout the samples, the metabolites are grouped hierarchically in the dendrogram on the left. Metabolites that behave similarly (with regard to abundance throughout the sample sites) are grouped together.

DISCUSSION

Discriminatory Power of Metabolomics

From the score plot, the groups are clearly divided along both components, indicating that the metabolomic profiles of the soil samples from each site are fairly different. This suggests that the divergence between the soil samples is probably being caused by differences in their microbial populations or chemical compositions. Site A is distinctly separated from the other groups, particularly along component 1, suggesting that the metabolites that contribute to component 1 are important in setting site A apart from the others. The soil samples from site A, which is the furthest apart from the other groups, come from an environment with a metabolomic signature that is noticeably distinct. This might be connected to particular microbial ecosystems, a lack of nutrients, or pollution.

Sites C and D show some separation along both Component 1 and Component 2, suggesting that while their metabolomic profiles are unique, they also share certain traits. For site D, the separation is more pronounced in the positive direction of component 1. Since site A is the group that is furthest from the other groups along this axis, component 1 (9.7%) is crucial in differentiating it from the others. The primary variations in soil chemistry or biology for site A are probably represented by the metabolites that contribute the most to Component 1. Component 2 (6.6%) aids in group differentiation, especially with regard to Groups B and C. The metabolites that contribute to Component 2 might be linked to minute variations in the microbial activity or soil composition among these groups.

The loading plot is shown in Figure 3, and the metabolites with the greatest impact on distinguishing the soil samples at different sites are those that are farthest from the origin (both positive and negative extremes) on the plot. These very significant metabolites may be associated with certain biological or environmental characteristics exclusive to particular sample sites, such as microbial activity, pollution, or healthy soil. The total variation is lessened by metabolites grouped close to the centre. These denote common metabolic processes that are present in all soil samples, irrespective of the location. It's critical to determine which metabolites are represented by these points in order to comprehend the loading plot's biological significance. lysine, prolyl-hydroxyproline (Pro-Hyp), acephate, phosphatidylserine, p-hydroxyhippuric acid, quercetin-3-O-beta-glucopyranosyl-6'-acetate, pyrithiobac, hetisine, bis(p-nitrophenyl) phosphate, arachidic

acid, etc. are a few of these important metabolites. Along with amino acid permease (AAP1 and AAP5), proline transporter (ProT2), and histidine transporters (LHT1 and LHT6), lysine is an amino acid transporter that has been found in plant roots [30, 31]. A variety of amino acids are taken up from the soil solution by these transporters, which differ in their substrate specificities and patterns of expression [30, 32]. By interacting with organic matter and microbial activity, lysine also contributes significantly to the nitrogen cycle in soil, albeit indirectly.

Microbes release lysine in the form of antibiotics [33], maybe as a biological warfare tactic to achieve supremacy over limited N sources in soil [34]. Microbes also employ lysine to synthesise osmoprotectants [35]. The presence of bis(p-nitrophenyl)phosphate, phosphatidylserine, and prolyl-hydroxyproline in soil is related to the decomposition of plant and animal matter, the use of fertilisers or pesticides that contain phosphate compounds, and the breakdown of organic matter.

The other important metabolites are pyrithiobac, a herbicide mostly used in agriculture to control broadleaf weeds; acephate, an organophosphate foliar and soil insecticide; and azoxystrobin, a systemic fungicide that is widely used in agriculture to control various fungal infections in crops. Exposure to these chemicals may cause cholinesterase inhibition, probable carcinogens, liver and kidney damage, and other health problems.

Top Contributing Metabolites

VIP

Prolyl-hydroxyproline, or Pro-Hyp, the metabolite with ID 1817, has the highest VIP score (~2.6), indicating that it is the most significant factor in differentiating soil samples at different sites. Lysine (1033) is the other prominent metabolite. Together, these metabolites contribute significantly to the explanation of sample variability. These top-ranking metabolites most likely have substantial discriminatory power, which aids in separating the various soil sites, as the PLS-DA model suggests. These metabolites may be related to the various sites' stress reactions, microbial activity, environmental conditions, or soil makeup. The heatmap displays variations in the abundance of metabolites at each of the four sampling sites (A, B, C, and D). Metabolite 1817, or Pro-Hyp, for instance, is far less common in some locations (shown in blue) and highly abundant in others (shown in

red). The possible ecological or environmental factors influencing soil metabolite levels at different sites can be better understood by looking at these abundance patterns.

Violin plot

In comparison to other groups, Site A displays a more widely distributed, broad, and somewhat symmetrical distribution around the mean. Though there is a noticeable tail that extends to both higher and lower concentrations, the density is minimum around the mean. This may indicate that the environmental or biological circumstances at site A varies, leading to a greater degree of variability in Pro-Hyp concentrations. Different soil properties (such as organic matter content and microbial diversity) or differing levels of microbial activity or collagen breakdown may have an impact on the samples' diversity.

Two peaks in the bimodal distribution of Site B signify two different subgroups inside the sampling point. Higher density is shown around these peaks in the larger centre portion, indicating that the majority of samples fall into two different concentration ranges. The bimodal pattern indicates that there might be two distinct sample clusters at site B, each with varying environmental factors affecting Pro-Hyp levels. These variations may reflect distinct microbial populations driving the synthesis or degradation of Pro-Hyp, or they may result from microenvironmental variations within the same sampling point.

In contrast to site B, site C has a more symmetrical, narrow distribution around the mean and less variability. This suggests that Group C's Pro-Hyp concentrations are more constant between samples. This stability could be the result of a more consistent biological mechanism or soil environment that powers Pro-Hyp metabolism. In this instance, the microbial communities engaged in Pro-Hyp dynamics may be more homogeneous, or there may be less environmental variation. Site D displays a distribution that is equally narrow, with a tail that extends towards lower concentrations. This implies that certain samples at site D have significantly lower quantities of Pro-Hyp than the majority, which have generally consistent concentrations. These lower concentrations can be the result of altered soil characteristics, decreased microbial activity, or slower collagen decomposition rates.

Collagen and other proline-rich proteins—which can come from plant roots, animal waste, or microbial activity in soil—are frequently broken down to produce Pro-Hyp. The variations in Pro-Hyp concentrations among the sampling sites may be the result of a number of ecological processes, including interactions between plants and soil, microbial activity, and soil organic matter.

Normalisation and VIP Score Importance

By normalising the concentrations, the normalisation procedure makes sure that they account for systematic variance, such as differences in sampling technique. Pro-Hyp is a crucial metabolite in differentiating across the sampling locations, as indicated by the VIP score. This could be because of its function as a marker of microbial activity in soils or its role in collagen metabolism. The observed variability among the groups suggests that Pro-Hyp may serve as a significant marker of the underlying ecological conditions or state of the soil.

SAM

The SAM plot indicates that 19 metabolites were found to be significant at the chosen delta value, as can be seen in the lower right panel. The 19 major metabolites include quercetin-3-O-beta-glucopyranosyl-6'-acetate, hetisine, arachidic acid, azoxystrobin (free acid), pro-Hyp, p-hydroxyhippuric acid, cryptochlorophaeic acid, l-lysine, acephate, phosphatidylserine, etc. With a False Discovery Rate (FDR) of 0.018, 1.8% of the significant discoveries could potentially be false positives. This is a very low error rate, suggesting strong confidence in the significance of these metabolites. It is expected that there are 0.67 false positives, which means that fewer metabolites are falsely identified as significant. According to the null hypothesis, 51.1% of metabolites should not be differently expressed, according to the p_0 value of 0.511. The cutoff values, cutlow: -Inf and cutup: 2.406, indicate the range of d(i) values above which metabolites are considered significant. There is a noticeable difference in the 19 significant metabolites between the various soil samples. These significant metabolites could indicate various soil conditions (nutrient availability, microbial activity, pollution) or metabolic processes occurring at various sampling sites.

Heat map

Based on their patterns throughout the samples, the metabolites are grouped hierarchically in the dendrogram on the left. Metabolites that behave similarly among the sample groups (in terms of abundance) are grouped together. These metabolites are more prevalent in sites A and B but less so in groups C and D, as indicated by the top cluster (rows 1417, 7111, etc.), which is predominantly red in groups A and B but blue in groups C and D. Acephate (ID 1417), the organophosphate insecticide, has been used extensively for decades to manage insect pests in agricultural settings. However, its use has been partially

banned in many countries because of its hazardous intermediate product methamidophos (Lin et al., 2020). Methamidophos is categorised as a class IV "highly toxic" pesticide, whereas acephate is a class II "moderately hazardous" pesticide. Because acephate is so soluble in water, it can quickly contaminate soil, groundwater, and plants. It can also be readily absorbed by plants and build up in their edible sections [36, 37].

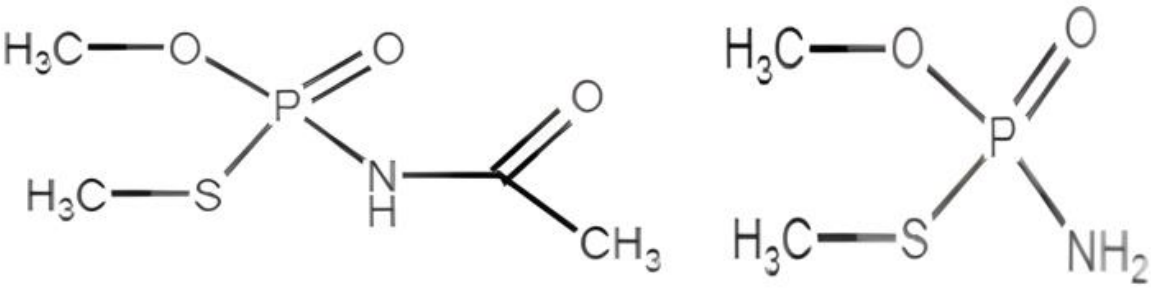

Figure 8: Structure of acephate and methamidophos (drawn with RCSB Protein Data Bank chemical sketch tool)

With the exception of C and D, all groups majorly display blue in the middle cluster (rows 3003, 5092, etc.), indicating that these metabolites are downregulated in sites A and B but abundant in sites C and D.

In conclusion, it appears that sites A and B have a wide variety of metabolite intensities, with some metabolites (upper portion) having high levels and others (middle and lower sections) having low levels. Site D exhibits variability, showing a balanced metabolomic profile with some metabolites substantially expressed and others suppressed.

CONCLUSION

The study found that the metabolomic profile of different soils, as determined by LC-MS, was able to discriminate soil samples from different sites. PLS-DA analysis showed distinct metabolomic differences between soil samples from different sites, with certain groups showing more similarities than others. Identification of key metabolites through the loading plot identifies the possible soil health or environmental stressors. The VIP plot highlighted important metabolites for distinguishing between soil samples, which could serve as biomarkers for specific soil conditions or biological processes.

The large number of identified yet unclassified metabolites found in our investigation highlights the present constraints concerning metabolite library sizes. The

composition of the unassigned metabolites may be important for discovering biomarkers or for expanding our present understanding of soil microbial function.

REFERENCES

1. Brown, R.W.; Chadwick, D.R.; Zang, H.; Jones, D.L. Use of Metabolomics to Quantify Changes in Soil Microbial Function in Response to Fertiliser Nitrogen Supply and Extreme Drought. Soil Biol. Biochem. 2021, 160, 108351. https://doi.org/10.1016/j.soilbio.2021.108351.
2. Chomel, M.; Guittonny-Larchevêque, M.; Fernandez, C.; Gallet, C.; DesRochers, A.; Paré, D.; Jackson, B.G.; Baldy, V. Plant Secondary Metabolites: A Key Driver of Litter Decomposition and Soil Nutrient Cycling. J. Ecol. 2016, 104, 1527–1541. https://doi.org/10.1111/1365-2745.12644.
3. Jones, O.A.H.; Sdepanian, S.; Lofts, S.; Svendsen, C.; Spurgeon, D.J.; Maguire, M.L.; Griffin, J.L. Metabolomic Analysis of Soil Communities Can Be Used for Pollution Assessment. Environ. Toxicol. Chem. 2014, 33, 61–64. https://doi.org/10.1002/etc.2418.
4. Lehmann, J.; Bossio, D.A.; Kögel-Knabner, I.; Rillig, M.C. The Concept and Future Prospects of Soil Health. Nat. Rev. Earth Environ. 2020, 1, 544–553. https://doi.org/10.1038/s43017-020-0080-8.
5. Pereira, P.; Bogunovic, I.; Muñoz-Rojas, M.; Brevik, E.C. Soil Ecosystem Services, Sustainability, Valuation and Management. Curr. Opin. Environ. Sci. Health 2018, 5, 7–13. https://doi.org/10.1016/j.coesh.2017.12.003.
6. Bünemann, E.K.; Bongiorno, G.; Bai, Z.; Creamer, R.E.; De Deyn, G.; de Goede, R.; Fleskens, L.; Geissen, V.; Kuyper, T.W.; Mäder, P.; et al. Soil Quality–A Critical Review. Soil Biol. Biochem. 2018, 120, 105–125. https://doi.org/10.1016/j.soilbio.2018.01.030.
7. Prăvălie, R. Exploring the Multiple Land Degradation Pathways across the Planet. Earth-Sci. Rev. 2021, 220, 103689. https://doi.org/10.1016/j.earscirev.2021.103689.
8. Borrelli, P.; Robinson, D.A.; Panagos, P.; Lugato, E.; Yang, J.E.; Alewell, C.; Wuepper, D.; Montanarella, L.; Ballabio, C. Land Use and Climate Change Impacts on Global Soil Erosion by Water (2015–2070). Proc. Natl. Acad. Sci. USA 2020, 117, 21994–22001. https://doi.org/10.1073/pnas.2001403117.
9. Kikuchi, J.; Ito, K.; Date, Y. Environmental Metabolomics with Data Science for Investigating Ecosystem Homeostasis. Prog. Nucl. Magn. Reson. Spectrosc. 2018, 104, 56–88. https://doi.org/10.1016/j.pnmrs.2017.11.003.
10. Matich, E.K.; Chavez Soria, N.G.; Aga, D.S.; Atilla-Gokcumen, G.E. Applications of Metabolomics in Assessing Ecological Effects of Emerging Contaminants and Pollutants on Plants. J. Hazard. Mater. 2019, 373, 527–535. https://doi.org/10.1016/j.jhazmat.2019.02.084.
11. Pétriacq, P.; Williams, A.; Cotton, A.; McFarlane, A.E.; Rolfe, S.A.; Ton, J. Metabolite Profiling of Non-Sterile Rhizosphere Soil. Plant J. 2017, 92, 147–162. https://doi.org/10.1111/tpj.13639.
12. van Dam, N.M.; Bouwmeester, H.J. Metabolomics in the Rhizosphere: Tapping into Belowground Chemical Communication. Trends Plant Sci. 2016, 21, 256–265. https://doi.org/10.1016/j.tplants.2016.01.008.

13. Bell, M.A.; McKim, U.; Sproule, A.; Tobalt, R.; Gregorich, E.; Overy, D.P. Extraction Methods for Untargeted Metabolomics Influence Enzymatic Activity in Diverse Soils. Sci. Total Environ. 2022, 828, 154433. https://doi.org/10.1016/j.scitotenv.2022.154433.
14. Swenson, T.L.; Jenkins, S.; Bowen, B.P.; Northen, T.R. Untargeted Soil Metabolomics Methods for Analysis of Extractable Organic Matter. Soil Biol. Biochem. 2015, 80, 189–198. https://doi.org/10.1016/j.soilbio.2014.10.007.
15. Simpson, M.J.; McKelvie, J.R. Environmental Metabolomics: New Insights into Earthworm Ecotoxicity and Contaminant Bioavailability in Soil. Anal. Bioanal. Chem. 2009, 394, 137–149. https://doi.org/10.1007/s00216-009-2612-4.
16. Rodríguez, A.; Castrejón-Godínez, M.L.; Salazar-Bustamante, E.; Gama-Martínez, Y.; Sánchez-Salinas, E.; Mussali-Galante, P.; Tovar-Sánchez, E.; Ortiz-Hernández, M.L. Omics Approaches to Pesticide Biodegradation. Curr. Microbiol. 2020, 77, 545–563. https://doi.org/10.1007/s00284-020-01916-5.
17. Patti, G.J.; Yanes, O.; Siuzdak, G. Innovation: Metabolomics: The Apogee of the Omics Trilogy. Nat. Rev. Mol. Cell Biol. 2012, 13, 263–269. https://doi.org/10.1038/nrm3314.
18. Wilson, I.; Plumb, R.; Granger, J.; Major, H.; Williams, R.; Lenz, E. HPLC-MS-Based Methods for the Study of Metabonomics. J. Chromatogr. B 2005, 817, 67–76. https://doi.org/10.1016/j.jchromb.2004.07.045.
19. Jones, O.A.H.; Maguire, M.L.; Griffin, J.L.; Dias, D.A.; Spurgeon, D.J.; Svendsen, C. Metabolomics and Its Use in Ecology. Aust. Ecol. 2013, 38, 713–720. https://doi.org/10.1111/aec.12019.
20. Tang, J. Microbial Metabolomics. Curr. Genom. 2011, 12, 391–403. https://doi.org/10.2174/138920211797248619.
21. Lehmann, J.; Hansel, C.M.; Kaiser, C.; Kleber, M.; Maher, K.; Manzoni, S.; Nunan, N.; Reichstein, M.; Schimel, J.P.; Torn, M.S.; et al. Persistence of Soil Organic Carbon Caused by Functional Complexity. Nat. Geosci. 2020, 13, 529–534. https://doi.org/10.1038/s41561-020-0612-3.
22. Fischer, H.; Meyer, A.; Fischer, K.; Kuzyakov, Y. Carbohydrate and Amino Acid Composition of Dissolved Organic Matter Leached from Soil. Soil Biol. Biochem. 2007, 39, 2926–2935. https://doi.org/10.1016/j.soilbio.2007.06.014.
23. Warren, C.R. Response of Osmolytes in Soil to Drying and Rewetting. Soil Biol. Biochem. 2014, 70, 22–32. https://doi.org/10.1016/j.soilbio.2013.12.008.
24. Kakumanu, M.L.; Cantrell, C.L.; Williams, M.A. Microbial Community Response to Varying Magnitudes of Desiccation in Soil: A Test of the Osmolyte Accumulation Hypothesis. Soil Biol. Biochem. 2013, 57, 644–653. https://doi.org/10.1016/j.soilbio.2012.08.014.
25. Roth, V.-N.; Dittmar, T.; Gaupp, R.; Gleixner, G. The Molecular Composition of Dissolved Organic Matter in Forest Soils as a Function of pH and Temperature. PLoS ONE 2015, 10, e0119188. https://doi.org/10.1371/journal.pone.0119188.
26. Baran, R.; Brodie, E.L.; Mayberry-Lewis, J.; Hummel, E.; Da Rocha, U.N.; Chakraborty, R.; Bowen, B.P.; Karaoz, U.; Cadillo-Quiroz, H.; Garcia-Pichel, F.; et al. Exometabolite Niche Partitioning among Sympatric Soil Bacteria. Nat. Commun. 2015, 6, 8289. https://doi.org/10.1038/ncomms9289.
27. 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.; et al. Proposed Minimum Reporting Standards for Chemical Analysis. Metabolomics 2007, 3, 211–221. https://doi.org/10.1007/s11306-007-0082-2.

28. Wellerdiek, M.; Winterhoff, D.; Reule, W.; Brandner, J.; Oldiges, M. Metabolic Quenching of Corynebacterium glutamicum: Efficiency of Methods and Impact of Cold Shock. Bioproc. Biosyst. Eng. 2009, 32, 581–592. https://doi.org/10.1007/s00449-008-0280-y.
29. Swenson, T.L.; Jenkins, S.; Bowen, B.P.; Northen, T.R. Untargeted Soil Metabolomics Methods for Analysis of Extractable Organic Matter. Soil Biol. Biochem. 2015, 80, 189–198. https://doi.org/10.1016/j.soilbio.2014.10.007.
30. Yao, X.; Nie, J.; Bai, R.; Sui, X. Amino Acid Transporters in Plants: Identification and Function. Plants 2020, 9, 972. https://doi.org/10.3390/plants9080972.
31. Li, F.; Dong, C.; Yang, T.; Bao, S.; Fang, W.; Lucas, W.J.; Zhang, Z. The Tea Plant CsLHT1 and CsLHT6 Transporters Take up Amino Acids, as a Nitrogen Source, from the Soil of Organic Tea Plantations. Hortic. Res. 2021, 8, 178. https://doi.org/10.1038/s41438-021-00615-x.
32. Feng, H.; Fan, X.; Miller, A.J.; Xu, G. Plant Nitrogen Uptake and Assimilation: Regulation of Cellular pH Homeostasis. J. Exp. Bot. 2020, 71, 4380–4392. https://doi.org/10.1093/jxb/eraa150.
33. Hamano, Y. Occurrence, Biosynthesis, Biodegradation, and Industrial and Medical Applications of a Naturally Occurring ε-Poly-L-Lysine. Biosci. Biotechnol. Biochem. 2011, 75, 1226–1233. https://doi.org/10.1271/bbb.110201.
34. Kielland, K. Landscape Patterns of Free Amino Acids in Arctic Tundra Soils. Biogeochemistry 1995, 31, 113–132. https://doi.org/10.1007/BF00000940.
35. Neshich, I.A.P.; Kiyota, E.; Arruda, P. Genome-Wide Analysis of Lysine Catabolism in Bacteria Reveals New Connections with Osmotic Stress Resistance. ISME J. 2013, 7, 2400–2410. https://doi.org/10.1038/ismej.2013.123.
36. Mohapatra, S.; Ahuja, A.K.; Deepa, M.; Sharma, D. Residues of Acephate and Its Metabolite Methamidophos in/on Mango Fruit (Mangifera indica L.). Bull. Environ. Contam. Toxicol. 2011, 86, 101–104. https://doi.org/10.1007/s00128-010-0154-2.
37. Syed, J.H.; Alamdar, A.; Mohammad, A.; Ahad, K.; Shabir, Z.; Ahmed, H.; Ali, S.M.; Sani, S.G.A.S.; Bokhari, H.; Gallagher, K.D.; et al. Pesticide Residues in Fruits and Vegetables from Pakistan: A Review of the Occurrence and Associated Human Health Risks. Environ. Sci. Pollut. Res. 2014, 21, 13367–13393. https://doi.org/10.1007/s11356-014-3117-z.