UNIVERSITA' DEGLI STUDI DI NAPOLI FEDERICO II

NATIONAL PH.D. PROGRAM IN ARTIFICIAL INTELLIGENCE

AGRIFOOD AND ENVIRONMENT

Thesis Defense for The Degree of Doctor of Philosophy

Artificial Intelligence approaches to explore Soil Microbiome Dynamics: predicting soil temperature from soil microbial community profiles with a Feed-Forward Artificial Neural Network (ANN)

Tutor:
Prof. Salvatore Cuomo

Co-tutor:
Michelina Ruocco

Coordinator:
Prof. Francesco Loreto

PhD Candidate:
Roberta Pace
Matr: DR996814

XXXVIII cycle - 2024/2025

Table of contents

i. Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favouring.

ii. Permission to use

The author grants permission for this thesis to be consulted and used exclusively for academic and research purposes. Any reproduction, distribution, or use of the material for commercial purposes, or any other purpose not related to scholarly work, is strictly prohibited without the author's prior written consent.

iii. Preface

This doctoral research was carried out within the National PhD program in Artificial Intelligence applied to Agrifood and Environment at the Institute for Sustainable Plant Protection (CNR-IPSP, Portici, Italy) and at the Evja s.r.l.. The work combined both experimental and computational approaches, part of the training and research activities was also conducted during a period abroad at the University of Salamanca (Spain), which provided the opportunity to broaden the scientific perspective and strengthen the methodological framework of the project.

My PhD thesis focuses on the study of the soil microbial communities in an agricultural system where four treatments were applied: (i) conventional farming without organic matter (C), (ii) conventional farming with the addition of organic matter (C+O), (iii) conventional farming with the addition of beneficial microorganisms but without organic matter (M), and (iv) conventional farming with the addition of both beneficial microorganisms and organic matter (M+O), to uncover microbial patterns across different treatments. Using metagenomic analyses (Oxford Nanopore Technology sequencing) and Artificial Intelligence (AI) approaches, I investigated how the community structure and evolves during the crop cycle. The results aim to improve the understanding of the mechanisms regulating biological soil fertility and to lay the foundation for developing predictive tools to support more sustainable agricultural management.

The whole research pathway was articulated over three years and followed a progressive structure. The first phase was devoted to the acquisition of specific skills in microbiology, bioinformatics, and statistical analysis, which allowed the design and refinement of the experimental workflow. The second phase focused on the collection of soil samples, sequencing of microbial communities, and the first round of bioinformatic and statistical analyses. The third phase integrated these datasets into advanced data analysis pipelines, applying machine learning (ML) methods to explore microbial community dynamics and implement an Artificial Neural Network (ANN) to predict soil temperature from microbial composition.

The outcomes of this work include a peer-reviewed publication, and a submitted original research article, together with additional analyses that complete the framework of the thesis. Beyond scientific outputs, the doctoral project also provided opportunities for collaboration, dissemination, and training, all of which contributed to shaping both the scientific and professional growth achieved during this period.

iv. Abstract of the thesis

Soil represents a fundamental component of terrestrial ecosystems and a cornerstone of agricultural productivity. Its health and fertility are largely governed by the complex interactions among microbial communities that drive nutrient cycling, organic matter turnover, and ecosystem resilience. However, understanding these interactions and their responses to agronomic management remains a major scientific challenge due to the high complexity and multidimensionality of microbiome data.

This PhD research, conducted within the National PhD in Artificial Intelligence (Agrifood and Environment area) at the Institute for Sustainable Plant Protection (CNR-IPSP, Portici, Italy), aimed to investigate soil microbial dynamics under different agronomic treatments by integrating metagenomic sequencing and AI-based analytical frameworks. Experimental work was carried out in a greenhouse cultivation of *Valerianella locusta* in Bagnolo Mella (Brescia, Northern Italy), where four treatments were applied: (i) conventional farming without organic matter (C), (ii) conventional farming with organic matter (C+O), (iii) conventional farming with beneficial microorganisms (M), and (iv) conventional farming with both microorganisms and organic matter (M+O). Soil samples collected at multiple time points were analysed using Oxford Nanopore sequencing (16S rRNA and ITS markers). Data processing and analysis were performed through Python-based pipelines, combining ML approaches to uncover patterns of microbial succession. Results revealed clear temporal-driven trajectories in both bacterial and fungal communities, and in some cases, with organic matter and microbial inoculation acting synergistically to enhance diversity and functional redundancy. Fungal communities showed higher sensitivity to treatments and temporal changes, while bacterial communities displayed greater stability and ecological resilience. The research outcomes include a review article published in *SN Applied Sciences* and an original research article submitted to *Physiologia Plantarum* for the Special Issue *"Smart Agriculture – BrIAS edition" (Brussels Institute for Advanced Studies)*, linked to the II Conference on Smart Agriculture (Brussels, 2025). Additional analyses developed within this thesis further integrated the results through ANN, reinforcing the predictive potential of microbial community data.

Ultimately, this work demonstrates how AI-driven approaches can bridge metagenomics and soil ecology, providing a methodological foundation for the development of predictive models of soil health. The results contribute to the broader goal of designing data-informed, sustainable agricultural systems where soil biodiversity is not only monitored but actively leveraged to ensure productivity, resilience, and environmental integrity.

Keywords: Artificial Neural Network; Machine Learning; Metagenomics; Soil microbiome; Sustainable Agriculture.

Chapter 1 – Introduction

1.1 Artificial Intelligence in Agrifood and Environmental Systems

Over the past two decades, AI has progressively emerged as a transformative tool in the agrifood and environmental domains (Nautiyal et al. 2025). Initially limited to exploratory applications in crop monitoring and resource optimization, AI has now become central to precision agriculture, environmental sustainability, and food system resilience (Getahun et al. 2024; Mustaza et al. 2025). The rapid increase in high-dimensional datasets – including genomic, metagenomic, climatic, and geospatial data – has accelerated the integration of ML and deep learning (DL) techniques into agricultural and environmental research (Benfenati et al. 2025; Botero-Valencia et al. 2025).

One of the major achievements of AI in this field is the development of predictive models, capable of identifying hidden patterns in complex datasets and forecasting crop performance, soil health, or the onset of biotic and abiotic stresses (Wilhelm et al. 2022). These models are being increasingly employed to optimize irrigation and fertilization (Jayarani et al. 2024; Umutoni and Samadi 2024; Adamo et al. 2025) detect plant diseases at early stages (Singla et al. 2024), and predict the long-term impact of climate change on agricultural productivity (Oikonomidis et al. 2023; Dhillon et al. 2024).

Looking ahead, the trends in AI applications for agrifood and the environment point toward an even greater integration of multilayered data sources (soil microbiome, remote sensing, climatic models, socio-economic data) and the adoption of explainable AI approaches to ensure transparency and reliability. Moreover, the transition from descriptive and correlative analyses toward prescriptive and predictive models marks a fundamental step in designing resilient and sustainable food systems. Ultimately, AI is not only reshaping how data are analysed, but also how decisions are made in the face of global challenges such as food security, soil degradation, and climate change.

1.2 The soil: role, fertility, and sustainability challenges

Soil represents one of the most complex and vital components of terrestrial ecosystems, acting as the foundation for agricultural productivity and a key regulator of global biogeochemical cycles. It provides a wide range of ecosystem services, including nutrient cycling, carbon sequestration, water retention and purification, and the maintenance of biodiversity (Meersmans et al. 2024). Beyond its role as a physical support for plants, soil is a living and dynamic system that sustains microbial, faunal, and plant communities, all of which interact to ensure ecological balance and resilience (Noronha et al. 2025).

Soil fertility is traditionally defined as the capacity of the soil to supply nutrients to plants in adequate amounts and proportions. However, modern concepts of fertility have evolved to include biological and ecological dimensions, encompassing microbial activity, organic matter turnover, and soil structure (Lehmann et al. 2020). These factors jointly determine the soil ability to support sustainable crop production over time. Biological fertility, in particular, has gained increasing attention as an indicator of soil health, linking microbial diversity and functional potential to nutrient cycling efficiency and plant productivity (Semenov et al. 2025). Despite its importance, soil is a non-renewable resource on a human timescale and is increasingly threatened by degradation processes such as erosion, compaction,

salinization, contamination, and loss of organic matter. Intensive agricultural practices, overuse of chemical inputs, and land-use changes have accelerated these processes, leading to a decline in soil health and ecosystem functionality (Kopittke et al. 2019). The loss of soil biodiversity further exacerbates these impacts, compromising resilience to environmental stress and reducing the ability to recover after disturbances (Tibbett et al. 2020; Eisenhauer et al. 2024). In recent years, regenerative and sustainable soil management approaches have emerged as promising strategies to counteract degradation and restore soil functionality. Practices such as reduced tillage, organic amendments, cover cropping, and microbial inoculation aim to rebuild soil organic matter and reactivate beneficial microbial networks (Khangura et al. 2023). Integrating these practices with data-driven monitoring systems allows for a more precise evaluation of soil health and a better understanding of the underlying ecological mechanisms (Huang et al. 2023). Moreover, preserving soil fertility is not only essential for maintaining agricultural productivity but also for achieving broader sustainability goals. Healthy soils are at the heart of climate regulation, food security, and biodiversity conservation, making their protection and regeneration a global priority for future generations. Healthy soils' ability to store and cycle carbon makes them one of the largest terrestrial carbon reservoirs, capable of mitigating greenhouse gas emissions and buffering the effects of climate change (Rodrigues et al. 2023). From an agricultural perspective, soil health is inseparable from food security. Productive and biologically active soils ensure the continuous availability of nutrients, water, and beneficial microbial interactions that support crop growth and resilience (Pandey and Saharan 2025). Degraded soils, conversely, reduce yields, lower nutritional quality, and increase dependence on external chemical inputs, threatening the sustainability of food production systems worldwide (Ghisman et al. 2025). Equally important is the role of soils as biodiversity hotspots. A single gram of soil can harbour billions of microorganisms—bacteria, fungi, archaea, and protists—alongside a wide range of invertebrates and plant roots, forming one of the most diverse and functionally complex ecosystems on Earth. This belowground biodiversity underpins ecosystem stability, resilience to perturbations, and the maintenance of aboveground life (Delgado-Baquerizo et al. 2025; Iqbal et al. 2025). Given these interconnected roles, the protection and regeneration of soils have become a global priority aligned with major international frameworks such as the United Nations Sustainable Development Goals (SDGs - https://sdgs.un.org/goals), particularly those addressing zero hunger (SDG 2), climate action (SDG 13), and life on land (SDG 15). Promoting soil conservation through sustainable management, monitoring, and restoration is therefore essential not only for agricultural productivity but also for ensuring planetary health and intergenerational equity.

1.3 The soil microbiome: composition, function, and ecological role

Beneath the surface, soils host an immense and dynamic biological network, composed of microorganisms that drive key ecological processes. Understanding the soil microbiome, its composition, diversity, and ecological functions, is therefore fundamental to advancing sustainable soil management and developing predictive models for ecosystem health. The soil microbiome encompasses bacteria, archaea, fungi, protists, and viruses, all interacting within a highly heterogeneous environment.

These microorganisms form intricate ecological networks that regulate key soil functions, nutrient cycling, organic matter decomposition, carbon storage, and plant–soil interactions (Creamer et al. 2022). Through their collective metabolic activities, they shape the chemical, physical, and biological properties of soils, directly influencing plant productivity and ecosystem resilience (Wang et al. 2021). Functionally, the soil microbiome acts as a biological engine sustaining soil fertility (Van Der Heijden et al. 2008). Microorganisms drive the transformation of organic residues into bioavailable forms of nitrogen, phosphorus, and other nutrients essential for plant growth (Creamer et al. 2022; Kiprotich et al. 2025). Symbiotic microorganisms, such as mycorrhizal fungi and nitrogen-fixing bacteria, establish mutualistic relationships that enhance plant nutrient uptake and stress tolerance. At the same time, antagonistic taxa play a crucial role in suppressing soil-borne pathogens, thus contributing to natural disease control and promoting plant health (Berendsen et al. 2012). From a compositional standpoint, the soil microbiome exhibits remarkable diversity both taxonomically and functionally. Bacteria and archaea are primarily responsible for nutrient mineralization and biogeochemical transformations, while fungi contribute to organic matter turnover, symbiotic associations, and soil aggregation (Kiprotich et al. 2025). The balance between these microbial groups, together with their spatial distribution across microhabitats, defines the functionality of the soil ecosystem (Fierer 2017). Importantly, microbial community composition is influenced by a multitude of factors, including soil type, vegetation, climate, land use, and agronomic management practices (Iqbal et al. 2025). Advances in metagenomic and multi-omics technologies have revolutionized the study of soil microbiome beyond the limits of traditional culture-based methods (Bertola et al. 2021). High-throughput sequencing, combined with bioinformatics and new technologies, now enables the exploration of microbial diversity, functional potential, and ecological interactions at an unprecedented scale (Thompson et al. 2019; McElhinney et al. 2022). These approaches are reshaping our understanding of how microbial communities respond to environmental gradients and management interventions, providing new opportunities for predictive modelling of soil health and ecosystem functionality. In summary, the soil microbiome is both a driver and indicator of soil health, reflecting the complex interplay between biological processes and environmental factors. Its study represents a cornerstone for developing sustainable agricultural systems, as it offers a window into the biological mechanisms that maintain soil fertility and ecosystem stability. Ultimately, the convergence of metagenomics and AI marks a new era in soil science, in which biological complexity can be decoded, predicted, and ultimately managed for sustainability.

1.4 Research objectives and experimental framework

Despite major advances in sequencing technologies and data analysis, the relationship between soil management, microbial community dynamics, and crop performance remains only partially understood. In particular, disentangling how different agronomic practices shape the biological fertility of soils requires integrative approaches that combine microbiological, environmental, and agronomic data.
Specifically, the objectives of the research were to:

- Characterize the taxonomic and functional composition of bacterial and fungal communities in soils subjected to different treatments combining organic matter addition and microbial inoculation.
- Monitor the temporal evolution of these communities throughout the crop cycle to identify successional patterns linked to plant development and soil management.
- Apply ML algorithms such as Principal Component Analysis (PCA), t-distributed stochastic neighbour embedding (t-SNE), and clustering analysis to detect hidden patterns.
- Apply a Feed-Forward ANN to predict soil temperature from microbiome composition.
- Develop an integrative analytical framework that could serve as the basis for future predictive tools supporting sustainable agricultural practices.

Through these objectives, the study aims to contribute both conceptually and methodologically to the emerging field of AI-assisted soil ecology, offering insights into how data-driven approaches can enhance our capacity to assess and manage soil health.

1.4.1 Study site description

The experimental trial was conducted in Bagnolo Mella (Brescia, Lombardy, Northern Italy) (Figure 1), within an agricultural site managed under conventional farming practices. The area is located in the Po Valley, one of the most intensively cultivated regions of Europe, characterized by a long agricultural tradition. The experimental field lies at approximately 45.46° N latitude and 10.13° E longitude, at an elevation of about 85 meters above sea level.

The region has a humid subtropical climate (Cfa) according to the Köppen–Geiger classification, with mild winters and warm summers. The mean annual temperature is around 13.5 °C, and the average annual rainfall is approximately 900 mm, mainly concentrated in spring and autumn. The site benefits from a temperate regime suitable for horticultural crops such as *Valerianella locusta* (lamb's lettuce), which was the model species adopted for this study.

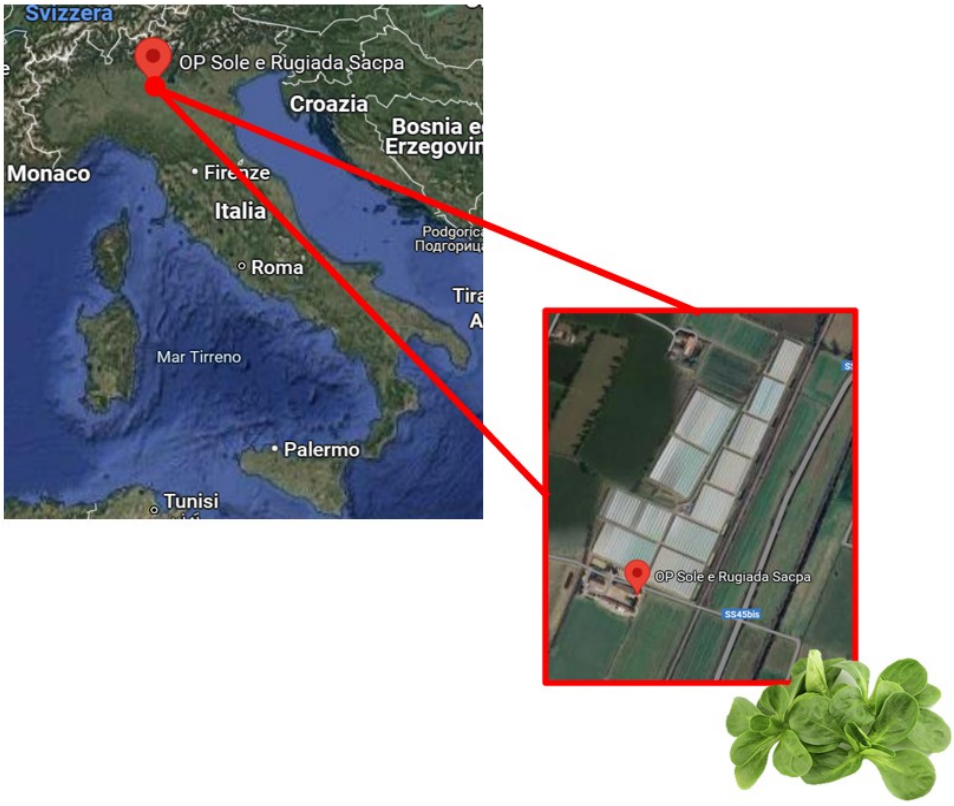


Figure 1. Bagnolo Mella farm location.

The soil of the experimental area is classified as a Typic Udifluvent, typical of the Po Plain alluvial deposits. The main physicochemical properties measured prior to the experiment (t_0 sampling), were as follows in Table 1:

Table 1. Physicochemical properties of the experimental soil before treatments application (Bagnolo Mella, Lombardy, Italy). Analytical results from Water & Life Lab S.r.l. (Report No. 22WL0084035, November 2022).

Parameter	Result	Unit	Interpretation
pH (H_2O)	7.55 ± 0.04	-	Neutral to slightly alkaline
pH (KCl)	6.8	-	-
Organic matter	1.55 ± 0.16	%	Low organic matter content
Electrical conductivity	697 ± 29.9	µS cm^{-1}	Moderate salinity
Cation exchange capacity (CEC, $BaCl_2$)	31.3 ± 0.18	meq 100 g^{-1}	High exchange capacity
Total nitrogen	1.9	g kg^{-1}	Low N content
C/N ratio	4.8	-	Low, limited organic N accumulation
Total $CaCO_3$	14.1 ± 0.27	g kg^{-1}	Moderately calcareous
Active $CaCO_3$	3.8 ± 0.13	g kg^{-1}	-
Exchangeable Ca	3782 ± 56	mg kg^{-1}	-
Exchangeable Mg	585 ± 32	mg kg^{-1}	-
Exchangeable K	257 ± 13	mg kg^{-1}	-
Exchangeable Na	249	mg kg^{-1}	-
Available P_2O_5 (Olsen)	223	mg kg^{-1}	High phosphorus availability
Cu (total)	32.3 ± 0.9	mg kg^{-1}	Within normal range
Sand (fine + coarse)	258	g kg^{-1}	-
Silt (fine + coarse)	592	g kg^{-1}	-
Clay	150	g kg^{-1}	-
Skeleton (>2 mm)	9.7	g kg^{-1}	Negligible

It is characterized by a silty loam texture, moderate drainage, and good structural stability. The relatively low organic matter and nitrogen contents highlight the need for organic and biological inputs, justifying the design of the four agronomic treatments tested in this study.

1.4.2 Experimental layout

The experiment was conducted within two multitunnel greenhouse structures, each divided in 2 plots, each plot covering a total surface area of 588 m^2 (8 m in width and 73.5 m in length). Sowing density was maintained constant across all treatments and crop cycles, corresponding to approximately 11 million seeds per hectare (≈1,100 seeds m^{-2}), which resulted in about 647,000 seeds per experimental plot (588 m^2).

More details on the experimental design can be found in the 'Materials and methods' section of Chapter 3.

This configuration allowed for adequate environmental control while maintaining realistic cultivation conditions representative of commercial production systems. The design also ensured sufficient replication to capture spatial variability within the greenhouse environment.

This operational schedule ensured consistency and repeatability across treatments, while allowing the monitoring of soil microbial dynamics in relation to key agronomic events. The repeated organic

amendments were planned to simulate realistic on-farm fertilization cycles and to evaluate their cumulative effects on soil microbial structure and fertility. Temporal sampling (t_0–t_5) was synchronized with the crop cycles.

1.4.3 Description of agronomic treatments

The experiment was designed to evaluate the combined and individual effects of organic matter addition and microbial inoculation on soil microbial communities and crop performance. Four treatments were applied as follows:

1. C (Control): Conventional farming without any organic or microbial input.
2. C+O: Conventional farming with the addition of organic matter.
3. M: Conventional farming with microbial inoculum but without organic matter.
4. M+O: Conventional farming with both microbial inoculum and organic matter.

The organic amendment consisted of well-composted bovine manure applied at a rate of 200 quintals per plot. Manure was incorporated into the top 15 cm of soil prior to sowing using light tillage.

Information on the crop cycle, the applications of organic matter and the average temperature ten days before harvesting are shown in Table 2-5.

Table 2. Information on the crop cycle of the control plot (conventional farming without addition of organic matter and of microorganisms).

C			
Sampling time	Sowing	Harvesting/Soil sampling	Average soil temperature (°C) 10 days before harvest
t0			6.50
t1	14/01/2023	20/03/2023	13.50
t2	12/04/2023	16/05/2023	18.54
t3	20/05/2023	15/06/2023	27.51
t4	26/07/2023	26/08/2023	26.79
t5	01/09/2023	05/10/2023	20.31

Table 3. Information on the crop cycle of the control + organic matter plot (conventional farming with addition of organic matter and without addition of microorganisms).

C+O				
Sampling time	Organic Matter addition	Sowing	Harvesting/Soil sampling	Average soil temperature (°C) 10 days before harvest
t0	✓			6.50
t1		14/01/2023	20/03/2023	13.50
t2		12/04/2023	16/05/2023	18.54
t3	✓	20/05/2023	21/06/2023	27.51
t4		26/07/2023	26/08/2023	26.79
t5	✓	01/09/2023	05/10/2023	20.31

Table 4. Information on the crop cycle of the plot treated with microorganisms (conventional farming without addition of organic matter and with addition of microorganisms).

M			
Sampling time	Sowing	Harvesting/Soil sampling	Average soil temperature (°C) 10 days before harvest
t0			6.5
t1	26/01/2023	29/03/2023	14.4
t2	12/04/2023	16/05/2023	18.5
t3	25/05/2023	22/06/2023	27.3
t4	01/07/2023	04/08/2023	24.0
t5	10/08/2023	13/09/2023	23.7

Table 5. Information on the crop cycle of the plot treated with microorganisms + organic matter (conventional farming with addition of organic matter and of microorganisms).

M+O				
Sampling time	Organic Matter addition	Sowing	Harvesting/ Soil sampling	Average soil temperature (°C) 10 days before harvest
t0	✓			6.5
t1		26/01/2023	28/03/2023	14.4
t2		12/04/2023	16/05/2023	18.5
t3	✓	25/05/2023	22/06/2023	27.3
t4		01/07/2023	04/08/2023	24.0
t5	✓	10/08/2023	13/09/2023	23.7

The microbial inoculum applied in the M and M+O plots was a commercial biostimulant "RyzoPepUp". Details on this product and its usage are in "Materials and methods" of the original research at Chapter 3.

All plots were managed under identical agronomic practices, irrigation, weeding, and pest control, ensuring that the only variables differentiating treatments were the organic and microbial inputs.

1.4.4 Environmental parameters

Environmental and agrometeorological conditions at the experimental site were monitored using EVJA smart farming devices (EVJA Srl) (Figure 2). The EVJA system integrates multiple sensors connected to a cloud-based data acquisition platform ("OPI"), enabling continuous (24-hour) monitoring of key microclimatic and pedoclimatic variables, including:

- Air temperature (−40 to 85 °C);
- Relative air humidity (0 to 100 % RH);
- Leaf wetness (0 to 100 %);
- Solar radiation (pyranometer; 0 to 250,000 lux / 0 to 2,000 W m^{-2});

- Soil water content (0 to 70 % VWC);
- Soil temperature (−40 to 60 °C);
- Soil electrical conductivity (0 to 20 dS m^{-1}).

Each microclimatic station was installed on mechanical supports specifically designed for protected cropping systems (Figure 3), with sensors positioned at canopy height to capture accurate data on the environmental factors influencing disease development and plant physiological responses.

The continuous data collected by the EVJA platform supported the farm management strategy, allowing the company to optimize irrigation, ventilation, and crop protection decisions throughout the production cycles.

For the purposes of this doctoral study, however, only a subset of the collected parameters was used: specifically, the mean soil temperature recorded during the 10 days preceding each harvest was extracted and correlated with the soil microbial community composition observed at the corresponding sampling time. This parameter was selected as a representative indicator of the short-term thermal regime affecting microbial activity and community structure at the soil–root interface.

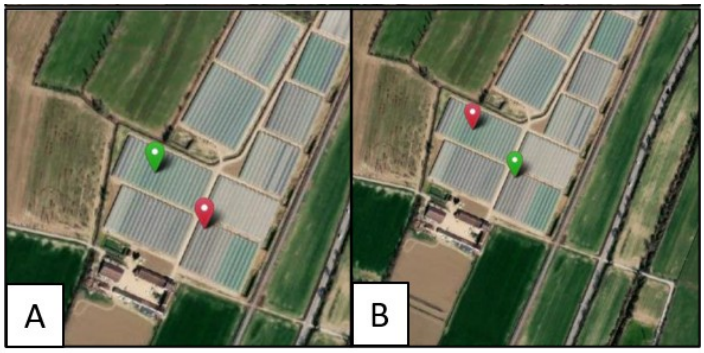


Figure 2. EVJA devices located in Bagnolo Mella (Brescia): in green (A) the device collecting data from the C and C+O plots (GPS coordinates 45.408084, 10.167536); (B) the device collecting data from the M and the M+O plots (GPS coordinates 45.407047, 10.168424).

Figure 3. EVJA device with data acquisition system "OPI".

1.5 References

Adamo T, Caivano D, Colizzi L, Dimauro G, Guerriero E (2025) Optimization of irrigation and fertigation in smart agriculture: an IoT-based micro-services framework. Smart Agricultural Technology 11: 100885

Benfenati D, Amalfitano D, Russo C, Tommasino C, Rinaldi AM (2025) Data centric artificial intelligence for the agrifood domain: a systematic mapping study. Comput Electron Agric 239: 110847

Berendsen RL, Pieterse CMJ, Bakker PAHM (2012) The rhizosphere microbiome and plant health. Trends Plant Sci 17: 478–486

Bertola M, Ferrarini A, Visioli G (2021) Improvement of soil microbial diversity through sustainable agricultural practices and its evaluation by -omics approaches: a perspective for the environment, food quality and human safety. Microorganisms 9: 1400

Botero-Valencia J, García-Pineda V, Valencia-Arias A, Valencia J, Reyes-Vera E, Mejia-Herrera M, Hernández-García R (2025) Machine learning in sustainable agriculture: systematic review and research perspectives. Agriculture 15: 377

Creamer RE, Barel JM, Bongiorno G, Zwetsloot MJ (2022) The life of soils: integrating the who and how of multifunctionality. Soil Biol Biochem 166: 108561

Delgado-Baquerizo M, Eldridge DJ, Liu YR, Liu ZW, Coleine C, Trivedi P (2025) Soil biodiversity and function under global change. PLoS Biol 23: e3003093

Dhillon R, Takoo G, Sharma V, Nagle M (2024) Utilizing a machine learning framework to evaluate the effect of climate change on maize and soybean yield. Comput Electron Agric 221: 108982

Eisenhauer N, Frank K, Weigelt A, Bartkowski B, Beugnon R, Liebal K, Mahecha M, Quaas M, Al-Halbouni D, Bastos A, Bohn FJ, de Brito MM, Denzler J, Feilhauer H, Fischer R, Fritsche I, Guimaraes-Steinicke C, Hänsel M, Haun DBM, Herrmann H, Huth A, Kalesse-Los H, Koetter M,

Kolleck N, Krause M, Kretschmer M, Leitão PJ, Masson T, Mora K, Müller B, Peng J, Pöhlker ML, Ratzke L, Reichstein M, Richter S, Rüger N, Sánchez-Parra B, Shadaydeh M, Sippel S, Tegen I, Thrän D, Umlauft J, Wendisch M, Wolf K, Wirth C, Zacher H, Zaehle S, Quaas J (2024) A belowground perspective on the nexus between biodiversity change, climate change, and human well-being. J Sustain Agric Environ 3: e212108

Fierer N (2017) Embracing the unknown: disentangling the complexities of the soil microbiome. Nat Rev Microbiol 15: 579–590

Getahun S, Kefale H, Gelaye Y (2024) Application of precision agriculture technologies for sustainable crop production and environmental sustainability: a systematic review. Scientific World Journal 2024: 2126734

Ghisman V, Muresan AC, Bogatu NL, Herbei EE, Buruiana DL (2025) Recent advances in the remediation of degraded and contaminated soils: a review of sustainable and applied strategies. Agronomy 15: 1920

Huang Y, Srivastava R, Ngo C, Gao J, Wu J, Chiao S (2023) Data-driven soil analysis and evaluation for smart farming using machine learning approaches. Agriculture 13: 1777

Iqbal S, Begum F, Nguchu BA, Claver UP, Shaw P (2025) The invisible architects: microbial communities and their transformative role in soil health and global climate changes. Environ Microbiome 20: 36

Jayarani P, Channa Basava P, Nagaraj P, Kusuma Latha P (2024) Optimizing fertilizer usage in agriculture with AI-driven recommendations. J Basic Sci Eng 21: 628–640

Khangura R, Ferris D, Wagg C, Bowyer J (2023) Regenerative agriculture—a literature review on the practices and mechanisms used to improve soil health. Sustainability 15: 2338

Kiprotich K, Muema E, Wekesa C, Ndombi T, Muoma J, Omayio D, Ochieno D, Motsi H, Mncedi S, Tarus J (2025) Unveiling the roles, mechanisms and prospects of soil microbial communities in sustainable agriculture. Discover Soil 2: 31

Kopittke PM, Menzies NW, Wang P, McKenna BA, Lombi E (2019) Soil and the intensification of agriculture for global food security. Environ Int 132: 105078

Lehmann J, Bossio DA, Kögel-Knabner I, Rillig MC (2020) The concept and future prospects of soil health. Nat Rev Earth Environ 1: 544–553

McElhinney JMWR, Catacutan MK, Mawart A, Hasan A, Dias J (2022) Interfacing machine learning and microbial omics: a promising means to address environmental challenges. Front Microbiol 13: 939461

Meersmans J, Colinet G, Negassa W (2024) Editorial: soil health, functions, and ecosystem services: insights into soil parameters and methods of integration. Front Environ Sci 12: 1358548

Mustaza SM, Mat Pauzi NA, Zainal N, Mohd Zaman MH, Mohamed Moubark A (2025) Artificial intelligence in precision agriculture: a review. Jurnal Kejuruteraan 37: 1025–1047

Nautiyal M, Joshi S, Hussain I, Rawat H, Joshi A, Saini A, Kapoor R, Verma H, Nautiyal A, Chikara A, Ahmad W, Kumar S (2025) Revolutionizing agriculture: a comprehensive review on artificial intelligence applications in enhancing properties of agricultural produce. Food Chem X 29: 102748

Noronha R, Sneppen K, Kaneko K (2025) Modelling soil as a living system: feedback between microbial activity and spatial structure. arXiv preprint arXiv:2502.20662

Oikonomidis A, Catal C, Kassahun A (2023) Deep learning for crop yield prediction: a systematic literature review. N Z J Crop Hortic Sci 51: 1–26

Pandey K, Saharan BS (2025) Soil microbiomes: a promising strategy for boosting crop yield and advancing sustainable agriculture. Discover Agriculture 3: 54

Rodrigues CID, Brito LM, Nunes LJR (2023) Soil carbon sequestration in the context of climate change mitigation: a review. Soil Syst 7: 64

Semenov MV, Zhelezova AD, Ksenofontova NA, Ivanova EA, Nikitin DA, Semenov VM (2025) Microbiological indicators for assessing the effects of agricultural practices on soil health: a review. Agronomy 15: 1763

Singla A, Nehra A, Joshi K, Kumar A, Tuteja N, Varshney RK, Gill SS, Gill R (2024) Exploration of machine learning approaches for automated crop disease detection. Curr Plant Biol 40: 100415

Thompson J, Johansen R, Dunbar J, Munsky B (2019) Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. PLoS One 14: e0215502

Tibbett M, Fraser TD, Duddigan S (2020) Identifying potential threats to soil biodiversity. PeerJ 8: e9271

Umutoni L, Samadi V (2024) Application of machine learning approaches in supporting irrigation decision making: a review. Agric Water Manag 294: 108710

Van Der Heijden MGA, Bardgett RD, van Straalen NM (2008) The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. Ecol Lett 11: 296–310

Wang C, Morrissey EM, Mau RL, Hayer M, Piñeiro J, Mack MC, Marks JC, Bell SL, Miller SN, Schwartz E, Dijkstra P, Koch BJ, Stone BW, Purcell AM, Blazewicz SJ, Hofmockel KS, Pett-Ridge J, Hungate BA (2021) The temperature sensitivity of soil: microbial biodiversity, growth, and carbon mineralization. ISME J 15: 2738–2747

Wilhelm RC, van Es HM, Buckley DH (2022) Predicting measures of soil health using the microbiome and supervised machine learning. Soil Biol Biochem 164: 108472

Chapter 2 – Review

2.1 Introduction

This chapter presents the review *"Artificial Intelligence in Soil Microbiome Analysis: a potential application in predicting and enhancing soil health - a review"*, published in *SN Applied Sciences* (Springer Nature). The article summarizes the state of the art in the application of AI to soil microbiome research, highlighting how ML and DL models are being increasingly used to analyse microbial diversity, predict soil health indicators, and model the impact of environmental and management factors. The review also discusses current limitations, such as data heterogeneity, lack of standardized methodologies, and the need for interpretability in AI models. Moreover, it identifies key research frontiers, including the integration of multi-omics datasets, the development of explainable AI, and the translation of predictive models into practical decision-support tools for sustainable agriculture.

By framing the soil microbiome at the intersection of biological complexity and computational innovation, this chapter provides the theoretical and methodological foundation for the experimental and analytical work presented in the following sections of this thesis. This review paper is reproduced in its published form, including the complete reference list and supplementary materials.

Discover Applied Sciences

Review

Artificial intelligence in soil microbiome analysis: a potential application in predicting and enhancing soil health—a review

Roberta Pace[1,2] · Vincenzo Schiano Di Cola[3] · Maurilia Maria Monti[2] · Antonio Affinito[4] · Salvatore Cuomo[3] · Francesco Loreto[1] · Michelina Ruocco[2]

Received: 26 August 2024 / Accepted: 25 November 2024
Published online: 16 January 2025
© The Author(s) 2024 OPEN

Abstract
Soil is a depletable and non-renewable resource essential for food production, crop growth, and supporting ecosystem services, such as the retaining and cycling of various elements, including water. Therefore characterization and preservation of soil biological health is a key point for the development of sustainable agriculture. We conducted a comprehensive review of the use of Artificial Intelligence (AI) techniques to develop forecasting models based on soil microbiota data able to monitor and predict soil health. We also investigated the potentiality of AI-based Decision Support Systems (DSSs) for improving the use of microorganisms to enhance soil health and fertility. While available studies are limited, potential applications of AI seem relevant to develop predictive models for soil fertility, based on its biological properties and activities, and implement sustainable precision agriculture, safeguarding ecosystems, bolstering soil resilience, and ensuring the production of high-quality food.

Article Highlights

- Maintaining healthy soil is crucial for sustainable agriculture to ensure long-term food production and environmental protection.
- Soil microbiota plays a main role in preserving soil health, therefore a correct analysis of soil microorganisms will help farmers understand and improve soil quality.
- Although research is limited, integrating AI technologies to monitor soil health holds great promise for sustainable precision agriculture.

Keywords Soil health · Microbial soil community · Artificial Intelligence · Forecasting models · Decision Support Systems

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42452-024-06381-4.

✉ Maurilia Maria Monti, mauriliamaria.monti@cnr.it | [1]Dipartimento di Biologia, Complesso Universitario Monte S. Angelo, Università degli Studi di Napoli Federico II, Edif.7, Via Cinthia, 80126 Napoli, Italia. [2]Istituto per la Protezione Sostenibile delle Piante, Consiglio Nazionale delle Ricerche, Piazzale Enrico Fermi, 1, 80055 Portici (Napoli), Italia. [3]Dipartimento di Matematica e Applicazioni R. Caccioppoli, Complesso Universitario Monte S. Angelo, Università degli Studi di Napoli Federico II, Via Cinthia, 80126 Napoli, Italia. [4]EVJA s.r.l., Via Benedetto Brin, 63, 80142 Napoli, Italia.

1 Introduction

Soil is a depletable and not renewable resource. Soil degradation, erosion, and contamination cause soil loss and can rapidly occur and significantly impact ecosystem services. Whereas, the formation of new layers of soil occurs at an often slower rate than soil loss, resulting in a net reduction of soil availability over time, and thus compromising agricultural productivity, biodiversity, water, and nutrient cycles [1, 2]. This highlights the importance of soil conservation and sustainable soil management practices, essential for maintaining soil biological health and ensuring its continued ability to support life on Earth [3].

The current systems of food production are among the main causes of the depletion of natural resources and environmental degradation. Notably, they affect the quality and resilience of the soil, making agriculture unsustainable [4, 5]. Even if soil quality and soil health are terms often used interchangeably [6], here we will use the term soil health. Soil quality is often associated with specific functions that soil performs, such as supporting plant growth, filtering water, and cycling nutrients. It focuses on the soil characteristics that contribute to these functions, assessing how well the soil can perform its roles in the ecosystem [7]. Soil quality is influenced by both natural and anthropogenic factors, including land management practices [8]. On the other hand, soil health is a broader concept that encompasses soil quality and is not just about current physical and chemical qualities but also about the soil overall fertility and sustainability over time [9].

Soil health recognizes that the soil is a living resource, harboring a diverse community of micro and macro-organisms [10–12]. These communities are part of the soil biodiversity and their differences in quality and quantity directly affect soil processes, such as the decomposition of organic matter, the solubilization of chemical elements, and the absorption and degradation of heavy metals and other pollutants [13, 14].

Intensive farming, monoculture, and limited crop rotations are the primary causes of loss of biodiversity, a main indicator of soil biological health [15, 16]. Sustainable land management practices, soil conservation techniques, and efforts to reduce soil pollution when farming are therefore crucial for preserving soil health and biodiversity and ensuring the long-term productivity of agriculture [17].

Nowadays, governments and international agencies are offering more support for sustainable agriculture through policies and research funding, recognizing the role of agriculture in achieving sustainability and development goals. For instance, under the EU's new policy framework "Food 2030," the European Commission advocates a novel development approach, aligned with the 17 Sustainable Development Goals of the United Nations, where sustainable land management is a pillar [18, 19].

Soil health also supports plant health [20–23] and this is the basis of the current EU "One Health" policies (https://health.ec.europa.eu/one-health_en). In agricultural soils, the dynamic interplay among biological, physical, and chemical properties is crucial. The microbial biomass plays a pivotal role in this process, influencing soil health, nutrient cycling, and overall agricultural productivity [24, 25].

In the recent past, agriculture has neglected soil multi-functionality and ecosystem services, leading to practices that have negatively impacted soil health [4]. Conventional agriculture, which uses chemical fertilization, pesticides, and intensive cultivation, can boost short-term productivity but damage soil by eroding, depleting nutrients, and reducing biodiversity [26, 27].

Although since 1970 the theoretical approach to agriculture has always had a focus on agronomic practices deemed sustainable, only from 2010 began to appear a better consciousness of the importance of agricultural management aimed at restoring ecosystem balance [28]. The intention now is to preserve soil fertility avoiding the use of synthetic agrochemicals, and minimizing the environmental impact of human activities with practices such as soft tillage, integrated management of the crops, and other environmentally friendly practices aimed to improve ecosystem services. These practices are preferred over conventional systems to maintain the optimum level of agroecosystem health status [8, 29–31]. Despite the growing attention on soil science, studying soils can be quite difficult due to their functional complexity [30–32], and finding the right indicators to assess whether the soil has suitable levels of quality and health is challenging [33–35].

The role of microorganisms in improving ecosystem soil health and promoting optimal plant development was largely discussed and deeply studied [36–41], and soil metagenome is now considered a "second genome" of plants [42]. The use of beneficial soil microorganisms, as biological pesticides or as biostimulants, for example, can be considered a sustainable agronomic practice to be implemented together with integrated management [43]. Augmentation of beneficial microorganisms to increase soil fertility can consequently be instrumental for improved crops and for other successful agricultural practices [44]. An ideal agricultural ecosystem should support microbial biodiversity with practices aimed at

maintaining local biological and ecological cycles increasing soil fertility, improving soil structure (e.g. by crop rotation), and preferring integrated pest management (IPM) with a limited use of chemicals [45–47].

1.1 Soil microbiota

Microbial soil communities have a fundamental role in establishing a network of positive interaction belowground [48, 49]. There are thousands of species of soil microorganisms that interact with each other and compete for nutrients, each of them with a specific role [50–52]. Through many of the enzymes and other substances they produce, microorganisms, especially bacteria and fungi, can stimulate plant growth, protect them from the attacks of pathogens and pests, and promote the mineralization of organic matter releasing nutrients necessary for plant development [53–55]. Many microorganisms form symbiotic associations with plants at the root level, playing an essential role in plant development by providing numerous benefits such as producing antibiotics, modulating the plant immune system, and acting as biostimulants [23, 56, 57], but the mechanisms that take place within the soil after the inoculum of such microorganisms is still unclear.

For hundreds of millions of years, a co-evolutionary phenomenon occurred, in which plants and microorganisms established strong symbiotic relationships [58] and influenced each other: plants can produce organic substances, especially under biotic or abiotic stresses, changing the surrounding chemical and physical environment, activating some groups of microorganisms, and removing others, and vice versa. Understanding and harnessing the complex interactions between soil, microorganisms and plants is essential for sustainable agriculture, ecosystem restoration, and environmental conservation efforts. It underscores the importance of maintaining healthy soil ecosystems for the well-being of agricultural systems. Using beneficial microbes, modern agriculture is not in contradiction with co-evolution: it is exploiting knowledge of the co-evolutionary phenomenon to optimize crop health and productivity under current conditions. We aim to integrate these principles, along with new technologies, to advance the development of management systems and soil treatment methods in open field conditions.

1.2 Methods for studying soil microorganisms

Microorganisms and their communities in soil can be assessed, both in terms of diversity and abundance, through different laboratory techniques: from the traditional techniques as culture on selective media to the most recent applications of multi-omics approaches [59], the latter largely improving precise and accurate overviews [7] and detection of microbial changes under different conditions and management practices, which may in turn affect soil properties [48]. Combining different techniques (the "multimethod" or "mixed-methods" approach) may also help cross-verify results, mitigating the limitations of individual methods, and gaining a more comprehensive understanding of the dynamics in the soil [60]. Although individual microorganisms are tiny and work within microenvironments, such as around a single root or within soil aggregates, their processes, such as nutrient cycling, organic matter breakdown, and pathogen suppression, affect larger, visible aspects of the environment like plant health, soil fertility, and carbon storage. We should better understand how to enhance soil bacterial processes, scaling them up from the microscale to levels that are significant for soil management. Sample size, along with replication, sampling design, DNA extraction, and molecular analysis methods, significantly influence measures of community structure under field conditions [61]. Scaling up results, from small sample size to represent a large field, requires careful consideration of these factors to avoid misrepresentation of microbial diversity. For this reason, it is important to consider standardized protocols like ISO 11063, which aim to optimize soil DNA extraction for reproducibility and comparability, though variations in methods and sample sizes can still impact the assessment of microbial abundance and diversity [62]. Also, modeling the dynamics of a microbial soil community can help understand belowground interactions that have repercussions on the soil ecosystem [63]. Among approaches based on soil total DNA analysis to provide taxonomic profiles of microbial communities, one of the most common methods is the metabarcoding: it is based on the amplification of specific DNA regions, such as 16S rDNA (ribosomal DNA) for bacteria and ITS (Internal Transcribed Spacer) for fungi, using PCR (Polymerase Chain Reaction), followed by high-throughput sequencing, e.g. with NGS (Next Generation Sequencing) methods [64]. Although this approach has the undeniable advantage of enabling the simultaneous analysis of many samples with a good level of precision and in a relatively short time, it does have some limitations that can affect the accuracy of taxonomic assignments. These drawbacks include potential biases introduced during DNA extraction and amplification [65], difficulties in distinguishing between live and dead organisms [66], resolution usually limited to family or genus level that does not permit differentiation between closely related species or sub-species [67], and challenges associated with incomplete

reference databases [65]. Additionally, the presence of PCR inhibitors in soil samples can complicate the process and lead to the underrepresentation of certain taxa. To address these limitations, researchers are exploring and developing alternative approaches and technologies. For example, to enhance taxonomic resolution, multiple genetic markers could be combined, and thanks to sequencing technologies developed by PacBio (Pacific Bioscience) and Oxford Nanopore, longer reads compared to traditional NGS methods are provided, enabling a better assembly of genomes and an accurate characterization of complex microbial communities [68]. To explore the genetic diversity and functional potential of complex microbial communities in a wide range of environments, metagenomics provides a powerful tool [69]. Instead of targeting specific genetic markers, Metagenomic Shotgun Sequencing involves the sequencing of all DNAs present in a sample providing a more comprehensive view, allowing the discovery of new species and functional genes.

Investigating the genera or the species of microorganisms present in soil can unveil valuable perspectives about the environment, as specific microorganisms can serve as "indicator species" or "bioindicators" [70, 71].

At a community level, the physiological profiling method performed using the Biolog® system is based on patterns of single carbon source utilization by microorganisms present in the soil. This analysis permits obtaining information on how microbial communities respond to changes of the environment over both spatial and temporal scales [72].

Enzymes produced by microorganisms are key factors in soil processes such as decomposing organic matter, cycling nutrients, and enhancing soil structure [73]. Measuring enzymatic activities of microorganisms in soil provides insights into microbial activity, and can serve as indirect indicators of soil health. Higher enzymatic activity levels often indicate greater microbial diversity, especially in the active soil microorganisms, essential for maintaining soil health [74–76]. An example of a soil sensor able to give information on biological enzyme activity was developed by DIGIT-soil [77]. This biosensor works through a customizable membrane targeting several soil enzymes, providing in-situ measurements and automated analysis of soil enzyme activity in few minutes. These sensors can be integrated into soil monitoring systems, allowing for continuous monitoring and timely interventions to maintain, restore, or improve soil fertility and health, providing real-time or near-real-time data on soil biological parameters. All the methods mentioned are summarized in Table 1.

1.3 Harnessing technologies for smarter agriculture

Agriculture 4.0 represented a significant transformation driven by the integration of cutting-edge technologies [85] to optimize crop yields while minimizing inputs such as water, fertilizer, and pesticides.

Big data analytics involves the examination of large and diverse datasets to identify hidden patterns, correlations, trends, and valuable insights, enabling farmers to make informed decisions based on a comprehensive understanding of factors influencing farming operations [86, 87]. Analysis of large datasets collected from sensors located proximally or distally from the cultivated area (e.g. drones or satellites), through the new computational technologies, can help farmers optimize resource allocation, predict crop yields, manage risks, and identify areas for improvement of farming practices [88].

Furthermore, devices, such as soil moisture sensors, weather stations, and automated irrigation systems, allow farmers to monitor and control agricultural processes in real-time. Connected sensors provide continuous data on environmental conditions and crop status, enabling precise management and timely interventions.

By integrating these innovations, farmers can advance beyond conventional methods, adopting a more data-driven and efficient approach that enables real-time informed decisions [89, 90]. Today, the decision-making process is facilitated by Decision Support Systems (DSSs) that gradually evolved to incorporate more sophisticated tools and techniques [91, 92]. With the incorporation of machine learning (ML) algorithms and predictive models, indeed, DSSs could analyze complex datasets, offering insights and recommendations for optimizing outcomes [93]. One of the most important questions in agriculture is to predict soil suitability for specific crops before planting [94]. In this context, a balanced microbiota supports optimal production yield while maintaining the ability of the soil to sustainably produce, which is a relevant entry of the DSS algorithm [95, 96], DSS capable of monitoring soil health must utilize many data inputs including soil nutrient levels, moisture content, temperature, microbial diversity, to feed ML algorithms aiming at predicting crop performances. DSSs integrating soil microbiology monitoring can also ensure that farmers can take action to restore microbial balance when necessary, by using beneficial microorganisms or compost amendments. This approach supports sustainable crop productivity and the maintenance and enrichment of soil health.

Data that allow DSS elaboration often need to be examined by Artificial Intelligence (AI). AI technologies, such as Machine Learning (ML) and Deep Learning (DL), have found numerous applications in agriculture, offering solutions to various challenges [97–103], and could result fundamental to analyze and model large amounts of data detected

Table 1 A glossary of the main methods mentioned in the paper used for studying soil microorganisms

Method	Basic principle	References
Metabarcoding	is a molecular biology technique used to identify and quantify species within a complex mixture of DNA, including environmental samples. It combines DNA barcoding, which involves sequencing a short, standardized region of DNA (a "barcode") that is unique to a species, with high-throughput sequencing (HTS) technologies, allowing for the simultaneous analysis of multiple species from a single sample	[78]
High-Throughput Sequencing (HTS)	refers to a set of advanced technologies that allow the rapid and large-scale sequencing of DNA or RNA. Unlike traditional Sanger sequencing, which sequences DNA one fragment at a time, HTS can sequence millions to billions of DNA fragments simultaneously, making it highly efficient and capable of generating vast amount of data in a relatively short time	[79]
Polymerase Chain Reaction (PCR)	is a widely used molecular biology technique that allows the amplification of specific DNA sequence, through key components: two primers, one forward and one reverse, short single-stranded DNA sequences that bind to opposite strands of the DNA template; an enzyme that synthesizes new DNA strands by adding nucleotides to the primers (polymerase); Deoxynucleotide Triphosphates (dNTPs), the building blocks (adenine [A], cytosine [C], guanine [G], and thymine [T]) used by DNA polymerase to synthesize the new DNA strands; a buffer solution that maintains the optimal pH and ionic conditions for the PCR reaction, ensuring that the DNA polymerase functions efficiently; the template, the DNA sequence that will be amplified. PCR involves three main steps: each cycle of denaturation, annealing, and extension doubles the amount of the target DNA sequence. After 20–40 cycles, the target DNA can be amplified millions to billions of times	[80]
Next Generation Sequencing (NGS)	is a collection of advanced sequencing technologies that enables the rapid and comprehensive sequencing of DNA or RNA. Unlike traditional Sanger sequencing, which sequences DNA one fragment at a time, NGS can sequence millions to billions of DNA fragments simultaneously, providing a massive amount of genetic information quickly and at a lower cost per base	[81]
PacBio sequencing	is an advanced DNA and RNA sequencing technology developed by Pacific Biosciences known as Single-Molecule Real-Time (SMRT) Sequencing. This technology is distinct from other sequencing methods due to its ability to produce extremely long reads, providing more comprehensive and accurate insights into genetic sequences, including complex regions that are difficult to sequence with other technologies	[81]
Oxford Nanopore sequencing	is an advanced DNA and RNA sequencing technology developed by Oxford Nanopore Technologies. This technology is known for its ability to produce ultra-long reads by directly sequencing nucleic acids as they pass through a nanopore, a tiny hole, embedded in a membrane. It is unique among sequencing technologies because it can generate real-time data, sequence very long fragments of DNA or RNA, and analyze a wide range of molecule types	[81]
Metagenomics	is the study of genetic material recovered directly from complex samples, rather than a single organism. This field allows scientists to analyze the collective genomes of entire communities of microorganisms, such as bacteria, archaea, viruses, and fungi, that inhabit a particular environment. The key advantage of metagenomics is that it enables the study of organisms that are difficult or impossible to culture in the lab, thus providing a more comprehensive understanding of microbial diversity, ecology, and function	[82]
Metagenomic Shotgun Sequencing	is a comprehensive and unbiased approach to study the genetic material within entire communities (including microbiota) directly from environmental samples. Unlike targeted sequencing methods, which focus on specific genes, shotgun metagenomics sequences all the DNA in a sample, capturing the full diversity of organisms and their functional potential	[83]
Biolog®	is a technology used for the identification and characterization of microorganisms based on their metabolic activities. It is commonly used in microbiology laboratories for the identification of bacteria, yeast, and fungi, as well as for assessing their metabolic profiles. It assesses the metabolic capabilities of microorganisms by measuring their ability to utilize various carbon sources and other substrates. It profiles different species or strains based on their metabolic fingerprints	[71]
Biosensors	are sophisticated analytical tools that integrate biological recognition elements with physical or chemical detectors to measure specific substances. Their applications span across medical diagnostics, environmental monitoring, food safety, biotechnology, and more, providing valuable insights and real-time data for various fields	[84]

in the field. AI-driven DSSs are instrumental for precision farming and its applications for sustainable agricultural practices [104], aimed at preserving ecosystems, increasing the resilience of soils, and ensuring the quality and quantity of food production (Agriculture 5.0) [105]. AI technologies are increasingly being utilized in agriculture to enhance productivity, efficiency, and sustainability of crops, and play a leading role in this context by enabling machines to perform tasks related to human intelligence [106]. Moreover, the integration of Internet of Things (IoT), Decision Support Systems (DSSs), Cloud Computing, Big data analytics and Remote Sensing has indeed brought significant advancements in agriculture [107, 108]. The complementation with automation and digitization systems enables both the collection of high-throughput data and the optimization of field-level practices such as planting [109–111], fertilization [112–114], irrigation [115–119], and pest control [120–122] based on real-time data acquisition and analysis, and often are also used for features such as data visualization, modeling, and simulation [123].

Numerous publications on generic applications of AI techniques for agricultural sustainability are available, which will not be reviewed here. Bhagat et al. (2022) made a bibliometric analysis, finding 465 papers only in the period 2000–2021 on this subject [124]. However, models developed using AI techniques to differentiate soil health status based on soil microbial composition, which serves as a fundamental input for sustainable agriculture, are scant in the literature.

1.4 AI empowers microbial community studies

For this review, we intended to consider only papers dealing with the prediction of soil health metrics through microbiome data as input, but this search was very restricting. So, we added studies that utilized AI techniques to predict outcomes based on microbiome composition, as well as those that employed AI to determine microbiome composition as an output. We consulted publications from "Web of Science", "SCOPUS" and "Google Scholar" databases until 1st October 2023, using the following keywords: "artificial intelligence", "soil microbial community", "soil health prediction", "soil microbiome monitoring", "soil health assessment", "soil microbiome", "soil microbial indicators", alone and using the Boolean operator "AND" to refine the research (e.g. "soil microbial community AND soil health prediction"), in all possible combination, excluding the redundant ones. Results are shown in Table S1. Out of 3508 papers found, only 12 encountered the requested criteria (Table 2), and their classification based on the AI approach applied is shown in Fig. 1.

As already mentioned, only a few studies have been published on how to predict soil health through information on microbial soil communities using AI techniques and forecasting models, despite the high potential of these applications [125]. In the context of studying soil microbiomes, researchers have proposed microbial-based metrics to evaluate soil health [76]. Indeed, Correa-Garcia et al. (2022) [125] claim that considering the ability of microorganisms to reflect past and present conditions of the habitats in which they are found, and are directly involved in the ecosystem processes, they also have some predictive capacity for future processes. Microbiomes are ideal for modeling and predicting the evolution of ecosystem processes as they participate directly in them. De Andrade et al. (2021) [105] propose guidelines for developing an Artificially Intelligent Soil Quality Index, incorporating input data for the ML model that includes recent land management practices, current genomic features, present and projected plant cover, as well as current soil biological, chemical, and physical parameters. Wilhelm et al. (2022) [126] aimed to develop a predictive model using soil bacterial community composition (in terms of 16S DNA sequences) to estimate 12 key soil health metrics, including biological, chemical, and physical properties. The study employed supervised machine learning algorithms, specifically Random Forest (RF) and Support Vector Machines (SVM), to classify and predict soil health metrics based on microbiome data collected from 949 soil samples from North American farmlands. Biological health metrics were the most accurately predicted, while the models struggled more with chemical and physical metrics, such as aggregate stability. However, the study suggests that larger and more region-specific datasets are needed to improve accuracy further. Hermans et al. (2020) [127] applied RF models to assess the degree to which bacterial communities in cultivated soils can be used to forecast physio-chemical soil characteristics: they verified a robust correlation between the composition of bacterial communities and land use, asserting that bacteria sensitivity to changes in specific soil parameters highlights the potential of bacterial DNA analysis as a tool for assessing soil quality. Chang et al. (2017) [128] applied a RF machine learning approach to predict crop productivity based on microbiome composition, revealing that differences in crop productivity were linked to the bulk soil microbiome composition and highlighting that several taxa are associated with nitrogen utilization.

Other publications were found on the use of AI techniques to study the relationship between soil microbial characteristics and the physical and chemical properties of soils [129, 130], to predict soil microbial biomass [131], to predict soil enzyme activity [132], and to assess soil microbiological characteristics or bacterial community structure [133–135].

Table 2 Reviewed papers

Research reviewed						
Ref	Year	Authors	Principle aspect	Category	Principle input	AI techniques applied
[105]	2021	de Andrade et al.	the authors propose guidelines for assembling an Artificially Intelligent Soil Quality Index, giving as input to the ML model data on recent land managements, current genomic features, current and projected plant-cover, time and current soil biological, chemical and physical parameters	Rewiev/Opinion paper	Microbiome data	None
[125]	2022	Correa-Garcia et al.	the authors discuss key criteria that differentiate the various statistical learning approaches and propose Supervised and Unsupervised learning methods to unlock forecasting potential of the microbiome for ecosystem processes	Rewiev/Opinion paper	Microbiome data	None
[126]	2022	Wilhelm et al.	the authors use 16S rRNA gene information to predict 12 measures of soil health through the application of supervised Machine Learning techniques	Original research paper	Microbiome data	Random Forest and Support Vector Machine regression and classification models
[127]	2020	Hermans et al.	the authors assess the degree to which bacterial communities in cultivated soils can forecast soil physico-chemical characteristics	Original research paper	Microbiome data	Random Forest model
[128]	2017	Chang et al.	the authors show that crop productivity differences were associated with bulk soil microbiome composition and highlighted several nitrogen utility-related taxa	Original research paper	Microbiome data	Random Forest model
[129]	2023	Sadeghi et al.	the authors used soil physical and chemical properties and total phospholipid fatty acid (PLFA) analysis to predict soil biological properties, and found a relationship between soil microbial characteristics and the physical and chemical properties of soil	Original research paper	Soil physical and chemical metrics	Cubist model algorithm
[130]	2017	Ebrahimi et al.	the authors evaluate the population of Azotobacter in soils with different land uses	Original research paper	Soil physical and chemical metrics	Artificial Neural Networks and Multivariate Linear Regression approaches

Table 2 (continued)

Research reviewed

Ref	Year	Authors	Principle aspect	Category	Principle input	AI techniques applied
[131]	2021	Pellegrini et al.	the authors predict soil microbial biomass from a range of physical and chemical properties	Original research paper	Soil biological physical and chemical metrics	Artificial Neural Network model approach
[132]	2012	Tajik et al.	the authors use soil and topographic attributes to predict the activity of three soil enzymes, L-asparaginase, L-glutaminase, and urease	Original research paper	Digital Terrain Analysis	Artificial Neural Networks and Multiple Linear Regression approaches
[133]	2012	Larsen et al.	the authors describe a bioclimatic modeling approach to predict microbial community structure as a function of environmental parameters and microbial interactions	Original research paper	Environmental parameters and microbial interactions	Artificial Neural Network model approach
[134]	2014	Santos et al.	the authors identify patterns in soil chemical variables and microbial community structures across forest soils in three regions of Brazil's Atlantic Forest	Original research paper	Soil physical and chemical metrics	Artificial Neural Network model approach
[135]	2018	Ramirez et al.	the authors merge 30 independent bacterial data sets comprising 1998 soil samples from 21 countries and show that disparate amplicon sequence data can be combined at the taxonomy-based level to assess bacterial community structure	Original research paper	Bacterial community	Random Forest models

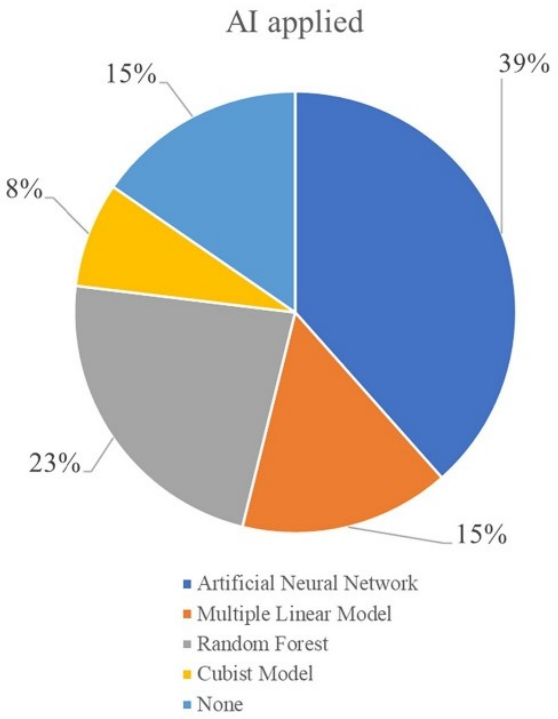


Fig. 1 Classification of the papers based on the AI techniques applied

All the above-mentioned papers are summarized in Table 2, with additional explainations of the main terms used in this paper in Table S2.

1.5 AI techniques applicable to soil microbiome analysis

The analysis of microbiome features can encompass a range of computational methods, from conventional statistical techniques to more sophisticated algorithms employed by AI. The first applications of AI to microbiota studies were used to investigate the role of human gut microbiota in human health: AI techniques such as ML and DL algorithms were applied for the diagnosis and therapy of critical diseases, with significant potential for improving public health [136–140]. An overview of ML methods for human microbiome analysis is given in Namkung (2020) [141].

Oh and Zhang (2020) [137] used autoencoders to convert microbiome data into low-dimensional representations, thereby enhancing the ability to predict human diseases. Curry et al. (2021) [142] also investigated ML techniques for disease detection from the gut microbiome, highlighting further computational and biological challenges. Morton et al. (2019) [143] presented a training process that aims to estimate the likelihood of detecting a metabolite given the provided microbe sequence through an Artificial Neural Network (ANN) approach. Tataru and David (2020) [144] used word-embedding techniques to analyze microbiome data and to better understand microbial interactions in human diseases. Costello and Martin (2018) [145] demonstrated how ML can predict metabolic pathways from time-series multi-omics data, thereby eliminating the need for traditional kinetic models. Metwally et al. (2019) [146] used longitudinal microbiome data to demonstrate the advantages of whole genome shotgun sequencing over amplicon sequencing, providing more insight into the microbiome's impact over time. Iadanza et al. (2020) [136] explored the use of AI in the identification and management of diseases by analyzing the gut microbiota. They emphasized the importance of AI in interpreting complex data sets. Liu et al. (2022) [147] created a ML framework to identify gut microbiome biomarkers across multiple geographic regions, utilizing ensemble learning and reinforcement learning techniques for personalized modulation analysis. Li et al. (2022) [148] investigated the use of multi-omics data to improve the accuracy of disease prediction models. They used ML techniques to study the gut microbiota and its effects on host metabolism. In summary, autoencoders and other ANN architectures have proven to be efficient in handling the complex and diverse

characteristics of microbiome data in medical science. This is critical for accurately predicting phenotypes and identifying biomarkers, as demonstrated by Hernández Medina et al. (2022) [149].

Expanding the use of ML techniques from medical to agricultural microbiome analysis presents numerous challenges, including data collection, sparsity, and biomarker identification. Soil samples differ significantly from controlled medical settings due to many factors, e.g. weather, topography, and host organism diversity. Furthermore, soil microbiome datasets are frequently smaller and less dense than human ones, thus needing more sophisticated data augmentation and handling techniques to avoid overfitting. While many studies on the human microbiome focus on disease prediction, the goal of agricultural applications often is the identification of biomarkers associated with soil fertility or crop yield. Nevertheless, the aforementioned medical examples can serve as a reference for applying AI techniques to study the soil microbiome. ML algorithms can significantly contribute to the advancement of our knowledge about soil health by analyzing data on microbial communities and making predictions about different environmental indicators [150].

The procedure for analyzing soil microbiome data using AI techniques adheres to a systematic sequence of steps commonly employed in bioinformatics and computational biology research (Fig. 2). Data pre-processing is an essential step in microbiome analysis to adequately prepare the input data for accurate ML and DL modeling. Normalization of microbiome data is necessary to adjust for differences in sequencing depths among samples. Various techniques such as rarefaction, total-sum scaling, or more advanced methods can be employed for this purpose [151]. To address skewness, logarithmic or other transformations such as square root can be employed. Missing values can be managed through techniques like k-nearest neighbors (k-NN), multiple imputation, or by assigning zero or minimal values, depending on the specific circumstances. In addition, when analyzing data that encompasses different taxa, creating new features might help in capturing more complex relationships that may not be apparent from the original data alone. Techniques such as Principal Component Analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) can be used to reduce the number of features and uncover structure in the data that may be related to the phenotype of interest. Dimensionality reduction is accomplished by using autoencoders to address the issue of high dimensionality in microbiome data. This method greatly accelerates model training and hyperparameter optimization [137]. After pre-processing, the next step is to choose and train an appropriate ML or DL model based on the hypothesis and nature of the data. Random Forests, Gradient Boosting Machines, and Support Vector Machines are common models for microbiome studies, as well as ANN architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Autoencoders for DL tasks. DL models, particularly CNNs and RNNs, are used to analyze complex microbiome datasets and predict diseases or phenotypic traits based on microbiome composition. These models are specifically intended to handle the massive amounts of data generated by high-throughput sequencing technologies [137]. Implementing cross-validation is fundamental because it helps ensure that the model does not just perform well on the training data but also generalizes well to new, unseen data, using techniques such as k-fold cross-validation, grid search, random search, or Bayesian optimization, that can help find the optimal model

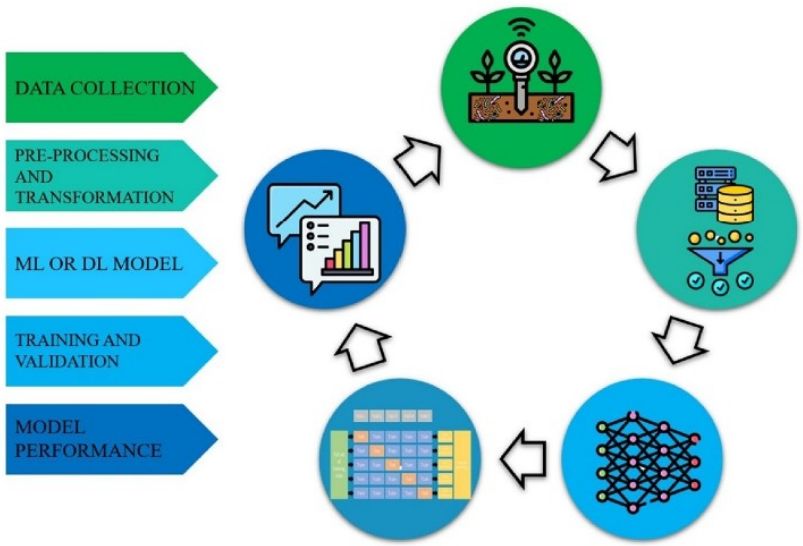


Fig. 2 Principal phases of developing a Machine Learning (ML) or Deep Learning (DL) model

settings [152]. Additionally, overfitting can be avoided by training the model on the training dataset and evaluating its performance on a validation set. The outcomes of the model evaluation, which occurs after it has been trained, can be interpreted within the framework of microbiome research. Metrics such as accuracy, precision, recall, F1-score for classification tasks, and mean squared error (MSE) or root mean squared error (RMSE) for regression tasks are computed to evaluate the performance of the model, depending on the specific task at hand. Models such as feature importance from tree-based models, SHAP values (Shapley Additive exPlanations), or permutation importance are among the commonly used to ascertain the taxa, features, or functions that are influencing the predictions of the model. By integrating diverse omics data, including metabolomics and metagenomics, and applying ML methodologies to this information, predictive models become more accurate. By adopting this workflow researchers can gain comprehensive insights into how microorganisms interact with one another and how these interactions influence their broader environment, whether this is a specific ecosystem, agricultural setting, or human body [137]. An overview of ML approaches useful for analyzing microbiome data is available in the paper by Papoutsoglou et al. (2023) [153]. A comprehensive review of utilizing DL techniques for analyzing environmental microbiome data is also provided by Deng et al. (2021) [154]. This last review outlines the advancements in DL techniques and elucidates their advantages over classical ML methods. The insights gained can be applied to predict outcomes or classify new samples in environmental settings. This workflow must be modified to reflect specific project goals, available data, and computational resources. Each step should be carefully planned and carried out to ensure the conclusions are reliable and valid.

A significant step forward would be the creation of AI-based forecasting models that integrate physical, chemical, and biological soil characteristics as well as agrometeorological data, allowing for the correlation of variation in microbial community composition with different levels of soil health. Most of the bases are in place to complete this step, but much work remains to be done on collecting data, soil parameter monitoring, parameter correlation, and mathematical model training for prediction. The initial step involves systematically collecting detailed information over time at regular intervals, focusing on soil properties such as pH, nutrient levels, organic matter content, and microbial composition. Scheduling and tailoring the samplings at consistent intervals (e.g., monthly, seasonally) to capture temporal variations, collecting them from multiple locations within the study area to account for spatial variability. This information could be obtained through classical soil sampling and analysis, and with the help of remote sensing technologies. Also, it is important to use a centralized database to store all collected data and record contextual information about sampling methods, locations, and conditions to provide context for data interpretation. Understanding the interactions between soil microorganisms and their surroundings is critical: the roles of various microbial species in soil health, their functions, and how they interact with plants can provide useful information. Using advanced molecular techniques such as metagenomics, researchers can identify beneficial microorganisms found in healthy soils, such as nitrogen-fixing bacteria, mycorrhizal fungi, and other symbiotic microbes that improve soil fertility and plant growth. Algorithms can be developed based on the collected data and understanding of microbial ecology to recommend specific microbial inoculants or management practices that are tailored to the soil's unique characteristics and the crops grown. By incorporating these elements into a DSS, farmers will be able to make more informed decisions that promote sustainable agriculture and productivity. Furthermore, farmers' feedback on the effectiveness of recommended practices and observed changes in soil health can be used to fine-tune the algorithms over time, which is critical for the DSS ongoing improvement. This entire process will take a long time to develop, but the results will be significant for farmers and the global agricultural sector.

The use of AI in microbiome research could revolutionize our understanding and control over microbial communities.

2 Concluding remarks

Globally, soils face severe degradation due to unsustainable agricultural practices, deforestation, and climate change impacts, leading to declines in soil fertility, erosion, and reductions in biodiversity (FAO on soil recarbonization practices: The Nature Conservancy https://www.nature.org/en-us).

The use of AI to predict soil health through microbiome information could represent a significant advancement in soil science. Microbiome data is inherently complex, with numerous microbial species interacting in intricate ways. Traditional analysis methods can be time-consuming and may overlook the full depth of these interactions. Therefore, microbiome data manipulation relies on key techniques to manage, analyze, and interpret the complex, high-dimensional datasets produced by sequencing technologies.

AI models can efficiently process vast amounts of data and identify patterns that may be difficult for humans to discern. Currently, the main examples of data manipulation on the microbiome are in medicine, while soil science

appears to be a new and still open field for AI applications. ML techniques, including RFs, SVMs, and ANNs, are widely used in microbial community classification, clustering, and prediction modeling. CNNs excel at interpreting microbial abundance data by transforming it into structured formats, whereas RNNs are especially effective at analyzing time-series data, such as modeling microbial succession dynamics. One significant advantage of AI models is their ability to constantly learn and adapt as new data becomes available. This adaptability is critical for studying soil microbiomes, which live in constantly changing environments. As new information becomes available, models can be refined to provide more accurate predictions of soil health. Furthermore, as discussed by Cuomo et al. (2024) [155], novel techniques such as Physics-Informed Neural Networks (PINNs) show promise in modeling the growth of microbial populations while accounting for environmental factors, providing a powerful tool for studying microbial dynamics.

Certainly, the advent of AI-based tools has the potential to significantly reduce the time and human resources needed. Future developments should prioritize collaboration between researchers, farmers, and industries for developing standardized datasets and protocols. This can improve the reliability and interoperability of AI models across different regions and agricultural systems, ensuring broader applicability and adoption. Understanding how well AI predicts soil health is essential for gaining the trust of farmers and other stakeholders, and is crucial for the widespread acceptance and adoption of AI-driven technologies.

In conclusion, the integration of AI for the prediction of soil health through microbiome information holds great promise for sustainable agriculture. Continued research, technological advancements, and collaborative efforts are essential to unlock the full potential of AI in optimizing soil management practices for a more resilient and productive agricultural future.

Acknowledgements This study was carried out within a National PhD in "Artificial Intelligence—Agrifood and Environment Area" co-funded by EVJA s.r.l.

Author contributions Conceptualisation: R.P., M.M.M. and M.R.; writing—original draft preparation: R.P., M.M.M. and M.R.; writing—review and editing: R.P., V.S.C., M.M.M., A.A., S.C., F.L. and M.R.; supervision: M.R.; funding acquisition: M.M.M. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding This study was carried out within the Agritech National Research Center and received funding from the European Union Next-Generation EU (Piano Nazionale di Ripresa e Resilienza (PNRR)—missione 4 componente 2, investimento 1.4—d.d. 1032 17/06/2022, cn00000022); and within the MISURA 16 "Cooperazione" SOTTOMISURA 16.2 – Tipologia di operazione 16.2.1 "Sostegno a progetti pilota e allo sviluppo di nuovi prodotti, pratiche, processi e tecnologie", CUP F59H23000010009. This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Montgomery DR. Soil erosion and agricultural sustainability. PNAS. 2007. https://doi.org/10.1073/pnas.0611508104.
2. Lal R. Restoring soil quality to mitigate soil degradation. Sustainability. 2015. https://doi.org/10.3390/su7055875.
3. Yin C, Zhao W, Pereira P. Soil conservation service underpins sustainable development goals. GECCO. 2022. https://doi.org/10.1016/j.gecco.2021.e01974.
4. Springmann M, Clark M, Mason-D'Croz D, Wiebe K, Bodirsky BL, Lassaletta L, de Vries W, Vermeulen SJ, Herrero M, Carlson KM, Jonell M, Troell M, DeClerck F, Gordon LJ, Zurayk R, Scarborough P, Rayner M, Loken B, Fanzo J, Charles HJG, Tilman D, Rockström J, Willett W. Options for keeping the food system within environmental limits. Nature. 2018. https://doi.org/10.1038/s41586-018-0594-0.

5. Giller KE, Delaune T, Silva JV, Descheemaeker K, van de Ven G, Schut AG, van Wijk M, Hammond J, Hochman Z, Taulya G, Chikowo R, Narayanan S, Kishore A, Bresciani F, Mancini Teixeira H, Andersson AJ, van Ittersum MK. The future of farming: who will produce our food? Food Secur. 2021. https://doi.org/10.1007/s12571-021-01184-6.
6. Bünemann EK, Bongiorno G, Bai Z, Creamer RE, De Deyn G, de Goede R, Fleskens L, Geissen V, Kuyper TW, Mäder P, Pulleman M, Sukkel W, Groenigen JW, Brussaard L. Soil quality–a critical review. Soil Biol Biochem. 2018. https://doi.org/10.1016/j.soilbio.2018.01.030.
7. Laishram J, Saxena KG, Maikhuri RK, Rao KS. Soil quality and soil health: a review. Int J Ecol Environ Sci. 2012;38(1):19–37.
8. Allen WJ, Bufford JL, Barnes DA, Barratt BIP, Deslippe JR, Dickie IA, Goldson SL, Howlett BG, Hulme PE, Lavorel S, O'Brien SA, Waller LP, Tylianakis JM. A network perspective for sustainable agroecosystems. Trends Plant Sci. 2022. https://doi.org/10.1016/j.tplants.2022.04.002.
9. Lehmann J, Bossio DA, Kögel-Knabner I, Rillig MC. The concept and future prospects of soil health. Nat Rev Earth Environ. 2020. https://doi.org/10.1038/s43017-020-0080-8.
10. Doran JW, Zeiss MR. Soil health and sustainability: managing the biotic component of soil quality. Appl Soil Eco. 2000. https://doi.org/10.1016/S0929-1393(00)00067-6.
11. Schloter M, Nannipieri P, Sørensen SJ, van Elsas JD. Microbial indicators for soil quality. Biol Fertil Soils. 2018. https://doi.org/10.1007/s00374-017-1248-3.
12. Ferreira CS, Seifollahi-Aghmiuni S, Destouni G, Ghajarnia N, Kalantari Z. Soil degradation in the European Mediterranean region: processes, status and consequences. Sci Total Environ. 2022. https://doi.org/10.1016/j.scitotenv.2021.150106.
13. Wagg C, Bender SF, Widmer F, Van Der Heijden MG. Soil biodiversity and soil community composition determine ecosystem multifunctionality. Proc Natl Acad Sci. 2014. https://doi.org/10.1073/pnas.1320054111.
14. Zhang J, Van Der Heijden MG, Zhang F, Bender SF. Soil biodiversity and crop diversification are vital components of healthy soils and agricultural sustainability. FASE. 2020. https://doi.org/10.5167/uzh-193808.
15. McLaughlin A, Mineau P. The impact of agricultural practices on biodiversity. Agric Ecosyst Environ. 1995. https://doi.org/10.1016/0167-8809(95)00609-V.
16. Bertola M, Ferrarini A, Visioli G. Improvement of soil microbial diversity through sustainable agricultural practices and its evaluation by-omics approaches. A perspective for the environment, food quality, and human safety. Microorganisms. 2021. https://doi.org/10.3390/microorganisms9071400.
17. Cárceles Rodríguez B, Durán-Zuazo VH, Soriano Rodríguez M, García-Tejero IF, Gálvez Ruiz B, Cuadros TS. Conservation agriculture as a sustainable system for soil health: a review. Soil Syst. 2022. https://doi.org/10.3390/soilsystems6040087.
18. European Commission and Directorate-General for Research and Innovation and Bizzo G, Fabbri K, Gajdzinska M, Haentjens W, Lueth D, Lutzeyer H, Luvarà G, Meloni G, Molina Villanueva J, Sainz Lopez N, Schoumacher C, Zampoukas N.Food 2030—Pathways for action 2.0—R&I policy as a driver for sustainable, healthy, climate resilient and inclusive food systems. Publications Office of the European Union. 2023; https://doi.org/10.2777/365011.
19. European Commission and European Research Council Executive Agency. Mapping ERC frontier research sustainable food production and consumption. Publications Office of the European Union. 2023; https://doi.org/10.2828/106806.
20. Ortiz A, Sansinenea E. The role of beneficial microorganisms in soil quality and plant health. Sustainability. 2022. https://doi.org/10.3390/su14095358.
21. Mohanty LK, Singh NK, Raj P, Prakash A, Tiwari AK, Singh V, Sachan P. Nurturing crops, enhancing soil health, and sustaining agricultural prosperity worldwide through agronomy. J Exp Agric Int. 2024. https://doi.org/10.9734/JEAI/2024/v46i22308.
22. Creamer RE, Brennan F, Fenton O, Healy MG, Lalor ST, Lanigan GJ, Regan JT, Griffiths BS. Implications of the proposed Soil Framework Directive on agricultural systems in Atlantic Europe–a review. Soil Use Manag. 2010. https://doi.org/10.1111/j.1475-2743.2010.00288.x.
23. Chaparro JM, Sheflin AM, Manter DK, Vivanco JM. Manipulating the soil microbiome to increase soil health and plant fertility. Biol Fertil Soils. 2012. https://doi.org/10.1007/s00374-012-0691-4.
24. Shahane AA, Shivay YS. Soil health and its improvement through novel agronomic and innovative approaches. Front Agron. 2021. https://doi.org/10.3389/fagro.2021.680456.
25. Kumar G, Baweja P, Gandhi PB. Impact of anthropogenic activities on soil patterns and diversity. Pedosphere. 2022. https://doi.org/10.1007/978-981-16-8770-9_13.
26. Norris CE, Congreves KA. Alternative management practices improve soil health indices in intensive vegetable cropping systems: a review. Front Environ Sci. 2018. https://doi.org/10.3389/fenvs.2018.00050.
27. Montoya D, Gaba S, de Mazancourt C, Bretagnolle V, Loreau M. Reconciling biodiversity conservation, food production, and farmers' demand in agricultural landscapes. Ecol Modell. 2020. https://doi.org/10.1016/j.ecolmodel.2019.108889.
28. Dönmez D, Isak MA, İzgü T, Şimşek Ö. Green horizons: navigating the future of agriculture through sustainable practices. Sustainability. 2024. https://doi.org/10.3390/su16083505.
29. Liu B, Tu C, Hu S, Gumpertz M, Ristaino JB. Effect of organic, sustainable, and conventional management strategies in grower fields on soil physical, chemical, and biological factors and the incidence of Southern blight. Appl Soil Ecol. 2007. https://doi.org/10.1016/j.apsoil.2007.06.007.
30. Rodrigo-Comino J, López-Vicente M, Kumar V, Rodríguez-Seijo A, Valkó O, Rojas C, Pourghasemi HR, Salvati L, Bakr N, Vaudour E, Brevik EC, Radziemska M, Pulido M, Di Prima S, Dondini M, de Vries W, Santos ES, de Lourdes M-S, Yu Y, Panagos P. Soil science challenges in a new era: a transdisciplinary overview of relevant topics. Air Soil Water Res. 2020. https://doi.org/10.1177/11786221209774.
31. Lehmann J, Hansel CM, Kaiser C, Kleber M, Maher K, Manzoni S, Nunan N, Reichstein M, Schimel JP, Torn MS, Wieder WR, Kögel-Knabner I. Persistence of soil organic carbon caused by functional complexity. Nat Geosci. 2020. https://doi.org/10.1038/s41561-020-0612-3.
32. Hou D, Bolan NS, Tsang DC, Kirkham MB, O'connor D. Sustainable soil use and management: an interdisciplinary and systematic approach. Sci Total Environ. 2020. https://doi.org/10.1016/j.scitotenv.2020.138961.
33. Rinot O, Levy GJ, Steinberger Y, Svoray T, Eshel G. Soil health assessment: a critical review of current methodologies and a proposed new approach. Sci Total Environ. 2019. https://doi.org/10.1016/j.scitotenv.2018.08.259.
34. Vogel HJ, Bartke S, Daedlow K, Helming K, Kögel-Knabner I, Lang B, Rabot E, Russell D, Stößel B, Weller U, Wiesmeier M, Wollschläger U. A systemic approach for modeling soil functions. 2018. Soil. https://doi.org/10.5194/soil-4-83-2018.

35. Evangelista SJ, Field DJ, McBratney AB, Minasny B, Ng W, Padarian J, Dobarco MR, Wadoux AMC. A proposal for the assessment of soil security: Soil functions, soil services and threats to soil. Soil Security. 2023. https://doi.org/10.1016/j.soisec.2023.100086.
36. Welbaum GE, Sturz AV, Dong Z, Nowak J. Managing soil microorganisms to improve productivity of agro-ecosystems. Crit Rev Plant Sci. 2004. https://doi.org/10.1080/07352680490433295.
37. de Salamone IEG. Use of soil microorganisms to improve plant growth and ecosystem sustainability. In: The molecular basis of plant genetic diversity. 2012. pp. 233–258.
38. Buscot F, Varma A. Microorganisms in soils: roles in genesis and functions. 2005; https://doi.org/10.1007/b137872.
39. Aislabie J, Deslippe JR, Dymond J. Soil microbes and their contribution to soil services. Ecosystem services in New Zealand–conditions and trends. 2013;1(12): 143–161.
40. Maron PA, Sarr A, Kaisermann A, Lévêque J, Mathieu O, Guigue J, Karimi B, Bernard L, Dequiedt S, Terrat S, Chabbi A, Ranjard L. High microbial diversity promotes soil ecosystem functioning. Appl Environ Microbiol. 2018. https://doi.org/10.1128/AEM.02738-17.
41. Aguilar-Paredes A, Valdés G, Nuti M. Ecosystem functions of microbial consortia in sustainable agriculture. Agronomy. 2020. https://doi.org/10.3390/agronomy10121902.
42. Banerjee S, van der Heijden MG. Soil microbiomes and one health. Nat Rev Microbiol. 2023. https://doi.org/10.1038/s41579-022-00779-w.
43. Hendriksen NB. Microbial biostimulants–the need for clarification in EU regulation. Trends Microbiol. 2022. https://doi.org/10.1016/j.tim.2022.01.008.
44. Bonanomi G, De Filippis F, Zotti M, Idbella M, Cesarano G, Al-Rowaily S, Abd-ElGawad A. Repeated applications of organic amendments promote beneficial microbiota, improve soil fertility and increase crop yield. Appl Soil Ecol. 2020. https://doi.org/10.1016/j.apsoil.2020.103714.
45. Ball BC, Bingham I, Rees RM, Watson CA, Litterick A. The role of crop rotations in determining soil structure and crop growth conditions. Can J Soil Sci. 2005. https://doi.org/10.4141/S04-078.
46. Hartmann M, Six J. Soil structure and microbiome functions in agroecosystems. Nat Rev Earth Environ. 2023. https://doi.org/10.1038/s43017-022-00366-w.
47. Trivedi P, Mattupalli C, Eversole K, Leach JE. Enabling sustainable agriculture through understanding and enhancement of microbiomes. New Phytol. 2021. https://doi.org/10.1111/nph.17319.
48. Dubey A, Malla MA, Khan F, Chowdhary K, Yadav S, Kumar A, Sharma S, Pramod K, Khare PK, Khan ML. Soil microbiome: a key player for conservation of soil health under changing climate. Biodivers Conserv. 2019. https://doi.org/10.1007/s10531-019-01760-5.
49. Jing J, Cong WF, Bezemer TM. Legacies at work: plant–soil–microbiome interactions underpinning agricultural sustainability. Trends Plant Sci. 2022. https://doi.org/10.1016/j.tplants.2022.05.007.
50. Bardget RD, Van Der Putten WH. Belowground biodiversity and ecosystem functioning. Nature. 2014. https://doi.org/10.1038/nature13855.
51. Bardgett RD, Caruso T. Soil microbial community responses to climate extremes: resistance, resilience, and transitions to alternative states. Philos T R Soc B. 2020. https://doi.org/10.1098/rstb.2019.0112.
52. Ali S, Moon YS, Hamayun M, Khan MA, Bibi K, Lee IJ. Pragmatic role of microbial plant biostimulants in abiotic stress relief in crop plants. J Plant Interact. 2022. https://doi.org/10.1080/17429145.2022.2091801.
53. Berendsen RL, Pieterse CM, Bakker PA. The rhizosphere microbiome and plant health. Trends Plant Sci. 2012. https://doi.org/10.1016/j.tplants.2012.04.001.
54. Tahat MM, Alananbeh KM, Othman YA, Leskovar DI. Soil health and sustainable agriculture. Sustainability. 2020. https://doi.org/10.3390/su12124859.
55. Iftikhar A, Farooq R, Akhtar M, Khalid H, Hussain N, Ali Q, Ul Malook S, Ali D. Ecological and sustainable implications of phosphorous-solubilizing microorganisms in soil. Discover Appl Sci. 2024. https://doi.org/10.1007/s42452-024-05683-x.
56. Balestrini R, Chitarra W, Fotopoulos V, Ruocco M. Potential role of beneficial soil microorganisms in plant tolerance to abiotic stress factors. In: Soil biological communities and ecosystem resilience. Springer International Publishing. 191–207, 2017; https://doi.org/10.1007/978-3-319-63336-7_12.
57. Caporale AG, Vitaglione P, Troise AD, Pigna M, Ruocco M. Influence of three different soil types on the interaction of two strains of Trichoderma harzianum with Brassica rapa subsp. sylvestris cv. esculenta, under soil mineral fertilization. Geoderma. 2019. https://doi.org/10.1016/j.geoderma.2019.05.003.
58. Wallenstein MD. Managing and manipulating the rhizosphere microbiome for plant health: a systems approach. Rhizosphere. 2017. https://doi.org/10.1016/j.rhisph.2017.04.004.
59. Enespa, Chandra P. Tool and techniques study to plant microbiome current understanding and future needs: an overview. Commun Integr Biol. 2022. https://doi.org/10.1080/19420889.2022.2082736.
60. Manfredini A, Malusà E, Costa C, Pallottino F, Mocali S, Pinzari F, Canfora L. Current methods, common practices, and perspectives in tracking and monitoring bioinoculants in soil. Front Microbiol. 2021. https://doi.org/10.3389/fmicb.2021.698491.
61. Philippot L, Ritz K, Pandard P, Hallin S, Martin-Laurent F. Standardisation of methods in soil microbiology: progress and challenges. FEMS Microbiol Ecol. 2012. https://doi.org/10.1111/j.1574-6941.2012.01436.x.
62. Penton CR, Gupta VV, Yu J, Tiedje JM. Size matters: assessing optimum soil sample size for fungal and bacterial community structure analyses using high throughput sequencing of rRNA gene amplicons. Front Microbiol. 2016. https://doi.org/10.3389/fmicb.2016.00824.
63. König S, Vogel HJ, Harms H, Worrich A. Physical, chemical and biological effects on soil bacterial dynamics in microscale models. Front Ecol Evol. 2020. https://doi.org/10.3389/fevo.2020.00053.
64. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, Thakare RP, Banday S, Alok K, Mishra AK, Das G, Malonia SK. Next-generation sequencing technology: current trends and advancements. Biology. 2023. https://doi.org/10.3390/biology12070997.
65. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol. 2012. https://doi.org/10.1111/j.1365-294X.2012.05470.x.
66. Wutkowska M, Vader A, Mundra S, Cooper EJ, Eidesen PB. Dead or alive; or does it really matter? Level of congruency between trophic modes in total and active fungal communities in high arctic soil. Front Microbiol. 2019. https://doi.org/10.3389/fmicb.2018.03243.

67. Rubiola S, Macori G, Civera T, Fanning S, Mitchell M, Chiesa F. Comparison between full-length 16S rRNA metabarcoding and whole metagenome sequencing suggests the use of either is suitable for large-scale microbiome studies. Foodborne Pathog Dis. 2022. https://doi.org/10.1089/fpd.2022.0027.
68. Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Front Genet. 2015. https://doi.org/10.3389/fgene.2015.00348.
69. Zapata JDD, Florez JEM, Alvarez DL. Metagenomics approaches to understanding soil health in environmental research-a review. Soil Sci. 2023. https://doi.org/10.37501/soilsa/163080.
70. Blagodatskaya E, Kuzyakov Y. Active microorganisms in soil: critical review of estimation criteria and approaches. Soil Biol Biochem. 2013. https://doi.org/10.1016/j.soilbio.2013.08.024.
71. Sofo A, Ricciuti P. A standardized method for estimating the functional diversity of soil bacterial community by Biolog® EcoPlatesTM assay—the case study of a sustainable olive orchard. Appl Sci. 2019. https://doi.org/10.3390/app9194035.
72. Ma F, Wang C, Zhang Y, Chen J, Xie R, Sun Z. Development of microbial indicators in ecological systems. Int J Environ Res Public Health. 2022. https://doi.org/10.3390/ijerph192113888.
73. Das SK, Varma A. Role of enzymes in maintaining soil health. Soil Biol. 2011. https://doi.org/10.1007/978-3-642-14225-3_2.
74. Klose S, Tabatabai MA. Urease activity of microbial biomass in soils as affected by cropping systems. Biol Fert Soils. 2000. https://doi.org/10.1007/s003740050645.
75. Ekenler M, Tabatabai M. β-Glucosaminidase activity of soils: effect of cropping systems and its relationship to nitrogen mineralization. Biol Fert Soils. 2022. https://doi.org/10.1007/s00374-002-0541-x.
76. Fierer N, Wood SA, de Mesquita CPB. How microbes can, and cannot, be used to assess soil health. Soil Biol Biochem. 2021. https://doi.org/10.1016/j.soilbio.2020.108111.
77. Iven H, Meller S, Luster J, Frossard E. A novel in-situ soil enzymatic activity sensor-expanding soil precision measurements to indicator of soil health. In EGU General Assembly Conference Abstracts. 2021; https://doi.org/10.5194/egusphere-egu21-10708.
78. Ruppert KM, Kline RJ, Rahman MS. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring, and applications of global eDNA. GECCO. 2019. https://doi.org/10.1016/j.gecco.2019.e00547.
79. Dalca AV, Brudno M. Genome variation discovery with high-throughput sequencing data. Brief Bioinform. 2010. https://doi.org/10.1093/bib/bbp058.
80. Dey P. Polymerase Chain Reaction: Principle, Technique and Applications in Pathology. In: Basic and Advanced Laboratory Techniques in Histopathology and Cytology. 2018; https://doi.org/10.1007/978-981-10-8252-8_20.
81. Marudamuthu B, Sharma T, Purru S, Soam SK, Rao CS. Next-generation sequencing technology: a boon to agriculture. Genet Resour Crop Evol. 2023. https://doi.org/10.1007/s10722-022-01512-5.
82. Nazir A. Review on metagenomics and its applications. Imp J Intersd Res. 2016;2:277–86.
83. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017. https://doi.org/10.1038/nbt.3935.
84. Ramesh M, Janani R, Deepa C, Rajeshkumar L. Nanotechnology-enabled biosensors: a review of fundamentals, design principles, materials, and applications. Biosens. 2022. https://doi.org/10.3390/bios13010040.
85. Javaid M, Haleem A, Singh RP, Suman R. Enhancing smart farming through the applications of Agriculture 4.0 technologies. Int J Intell Netw. 2022. https://doi.org/10.1016/j.ijin.2022.09.004.
86. Ang KLM, Seng JKP. Big data and machine learning with hyperspectral information in agriculture. IEEE Access. 2021. https://doi.org/10.1109/ACCESS.2021.3051196.
87. Bhat SA, Huang NF. Big data and AI revolution in precision agriculture: survey and challenges. IEEE Access. 2021. https://doi.org/10.1109/ACCESS.2021.3102227.
88. Cravero A, Pardo S, Sepúlveda S, Muñoz L. Challenges to use machine learning in agricultural big data: a systematic literature review. Agronomy. 2022. https://doi.org/10.3390/agronomy12030748.
89. Purnama S, Sejati W. Internet of things, big data, and artificial intelligence in the food and agriculture sector. Int Trans Artif Intell. 2023. https://doi.org/10.33050/italic.v1i2.274.
90. Oliveira LF, Moreira AP, Silva MF. Advances in agriculture robotics: a state-of-the-art review and challenges ahead. Robotics. 2021. https://doi.org/10.3390/robotics10020052.
91. Talari G, Cummins E, McNamara C, O'Brien J. State of the art review of Big Data and web-based Decision Support Systems (DSS) for food safety risk assessment with respect to climate change. Trends Food Sci Technol. 2022. https://doi.org/10.1016/j.tifs.2021.08.032.
92. Fenu G, Malloci FM. DSS LANDS: a decision support system for agriculture in Sardinia. HighTech Innov J. 2020. https://doi.org/10.28991/HIJ-2020-01-03-05.
93. Folorunso O, Ojo O, Busari M, Adebayo M, Joshua A, Folorunso D, Okechukwu Ugwunna C, Olabanjo O, Olabanjo O. Exploring machine learning models for soil nutrient properties prediction: a systematic review. BDCC. 2023. https://doi.org/10.3390/bdcc7020113.
94. Araújo SO, Peres RS, Ramalho JC, Lidon F, Barata J. Machine learning applications in agriculture: current trends, challenges, and future perspectives. Agronomy. 2023. https://doi.org/10.3390/agronomy13122976.
95. Smercina DN, Bailey VL, Hofmockel KS. Micro on a macroscale: relating microbial-scale soil processes to global ecosystem function. FEMS Microbiol Ecol. 2021. https://doi.org/10.1093/femsec/fiab091.
96. Wagg C, Hautier Y, Pellkofer S, Banerjee S, Schmid B, van der Heijden MG. Diversity and asynchrony in soil microbial communities stabilizes ecosystem functioning. Elife. 2021. https://doi.org/10.7554/eLife.62813.
97. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: a review. Sensors. 2018. https://doi.org/10.3390/s18082674.
98. Kamilaris A, Prenafeta-Boldú FX. Deep learning in agriculture: a survey. Comput Electron Agr. 2018. https://doi.org/10.1016/j.compag.2018.02.016.
99. Arvanitis KG, Symeonaki EG. Agriculture 4.0: the role of innovative smart technologies towards sustainable farm management. Open Agric. 2020. https://doi.org/10.2174/1874331502014010130.

100. Ren C, Kim DK, Jeong D. A survey of deep learning in agriculture: techniques and their applications. J Inf Process Syst. 2020. https://doi.org/10.3745/JIPS.04.0187.
101. Santos L, Santos FN, Oliveira PM, Shinde P. Deep learning applications in agriculture: a short review. Robot. 2020. https://doi.org/10.1007/978-3-030-35990-4_12.
102. Benos L, Tagarakis AC, Dolias G, Berruto R, Kateris D, Bochtis D. Machine learning in agriculture: a comprehensive updated review. Sensors. 2021. https://doi.org/10.3390/s21113758.
103. Meshram V, Patil K, Meshram V, Hanchate D, Ramkteke SD. Machine learning in agriculture domain: a state-of-art survey. Artif Intell Life. 2021. https://doi.org/10.1016/j.ailsci.2021.100010.
104. Naman A, Tiyndel GJ, Bhatia JK, Bhardwaj N, Rahul. Utilizing artificial intelligence (AI) for sustainable agriculture: precision farming as a catalyst for environmental conservation. Int J Agric Extension Soc Dev. 2024. https://doi.org/10.33545/26180723.2024.v7.i3e.441.
105. de Andrade VHGZ, Redmile-Gordon M, Barbosa BHG, Andreote FD, Roesch LFW, Pylro VS. Artificially intelligent soil quality and health indices for 'next generation' food production systems. Trends Food Sci. 2021. https://doi.org/10.1016/j.tifs.2020.10.018.
106. Camaréna S. Artificial intelligence in the design of the transitions to sustainable food systems. J Clean Prod. 2020. https://doi.org/10.1016/j.jclepro.2020.122574.
107. Bharman P, Saad SA, Khan S, Jahan I, Ray M, Biswas M. Deep learning in agriculture: a review. Asian J Res Comput Sci. 2022. https://doi.org/10.9734/AJRCOS/2022/v13i230311.
108. Alahmad T, Neményi M, Nyéki A. Applying IoT sensors and big data to improve precision crop production: a review. Agronomy. 2023. https://doi.org/10.3390/agronomy13102603.
109. Raj R, Aravind A, Akshay VS, Chandy M, Sharun ND. A seed planting robot with two control variables. ICOEI. 2019. https://doi.org/10.1109/ICOEI.2019.8862541.
110. Srinivasan N, Prabhu P, Smruthi SS, Sivaraman NV, Gladwin SJ, Rajavel R, Natarajan AR. Design of an autonomous seed planting robot. IEEE R10-HTC. 2016. https://doi.org/10.1109/R10-HTC.2016.7906789.
111. Ku HH, Liu CH, Wang WC. Design of an artificial intelligence of things based indoor planting model for mentha spicata. Processes. 2022. https://doi.org/10.3390/pr10010116.
112. Fiorentini M, Schillaci C, Denora M, Zenobi S, Deligios PA, Santilocchi R, Perniola M, Ledda L, Orsini R. Fertilization and soil management machine learning based sustainable agronomic prescriptions for durum wheat in Italy. Precis Agric. 2024. https://doi.org/10.1007/s11119-024-10153-w.
113. Gao J, Zeng W, Ren Z, Ao C, Lei G, Gaiser T, Srivastava AK. A fertilization decision model for maize, rice, and soybean based on machine learning and swarm intelligent search algorithms. Agronomy. 2023. https://doi.org/10.3390/agronomy13051400.
114. Firmansyah E, Pardamean B, Ginting C, Mawandha HG, Putra DP, Suparyanto T. Development of artificial intelligence for variable rate application based oil palm fertilization recommendation system. ICIMTech. 2021. https://doi.org/10.1109/ICIMTech53080.2021.9535082.
115. Talaviya T, Shah D, Patel N, Yagnik H, Shah M. Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. Artif Intell Agric. 2020. https://doi.org/10.1016/j.aiia.2020.04.002.
116. Tace Y, Elfilali S, Tabaa M, Leghris C. Implementation of smart irrigation using IoT and artificial intelligence. Math Model Comput. 2023. https://doi.org/10.23939/MMC2023.02.575.
117. Benhmad T, Rhaimi CB, Alomari S, Aljuhani L. Design and implementation of an integrated iot and artificial intelligence system for smart irrigation management. Int J Adv Soft Comput Appl. 2024. https://doi.org/10.15849/IJASCA.240330.12.
118. Biswas S, Sharma LK, Ranjan R, Saha S, Chakraborty A, Banerjee JS. Smart farming and water saving-based intelligent irrigation system implementation using the internet of things. In Recent trends in computational intelligence enabled research. 2021; https://doi.org/10.1016/B978-0-12-822844-9.00043-8.
119. Klyushin D, Tymoshenko A. Optimization of drip irrigation systems using artificial intelligence methods for sustainable agriculture and environment. Artificial intelligence for sustainable development: Theory, practice and future applications. 2021; https://doi.org/10.1007/978-3-030-51920-9_1.
120. Bhoi SK, Jena KK, Panda SK, Long HV, Kumar R, Subbulakshmi P, Jebreen HB. An Internet of Things assisted Unmanned Aerial Vehicle based artificial intelligence model for rice pest detection. Microprocess Microsyst. 2021. https://doi.org/10.1016/j.micpro.2020.103607.
121. Balaska V, Adamidou Z, Vryzas Z, Gasteratos A. Sustainable crop protection via robotics and artificial intelligence solutions. Machines. 2023. https://doi.org/10.3390/machines11080774.
122. Li C, Wang M. Pest and disease management in agricultural production with artificial intelligence: innovative applications and development trends. Adv Res. 2024. https://doi.org/10.50908/arr.4.3_381.
123. Zhai Z, Martínez JF, Beltran V, Martínez NL. Decision support systems for agriculture 4.0: survey and challenges. Comput Electron Agric. 2020. https://doi.org/10.2174/1874331502014010130.
124. Bhagat PR, Naz F, Magda R. Artificial intelligence solutions enabling sustainable agriculture: a bibliometric analysis. PLoS ONE. 2022. https://doi.org/10.1371/journal.pone.0268989.
125. Correa-Garcia S, Constant P, Yergeau E. The forecasting power of the microbiome. Trends Microbiol. 2022. https://doi.org/10.1016/j.tim.2022.11.013.
126. Wilhelm RC, van Es HM, Buckley DH. Predicting measures of soil health using the microbiome and supervised machine learning. Soil Biol Biochem. 2022. https://doi.org/10.1016/j.soilbio.2021.108472.
127. Hermans SM, Buckley HL, Case BS, Curran-Cournane F, Taylor M, Lear G. Using soil bacterial communities to predict physico-chemical variables and soil quality. Microbiome. 2020. https://doi.org/10.1186/s40168-020-00858-1.
128. Chang HX, Haudenshield JS, Bowen CR, Hartman GL. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. Front Microbiol. 2017. https://doi.org/10.3389/fmicb.2017.00519.
129. Sadeghi S, Petermann BJ, Steffan J, Brevik EC, Gedeon C. Predicting microbial responses to changes in soil physical and chemical properties under different land management. Appl Soil Ecol. 2023. https://doi.org/10.1016/j.apsoil.2023.104878.
130. Pellegrini E, Rovere N, Zaninotti S, Franco I, De Nobili M, Contin M. Artificial neural network (ANN) modelling for the estimation of soil microbial biomass in vineyard soils. Biol Fertil Soils. 2021. https://doi.org/10.1007/s00374-020-01498-1.

131. Tajik S, Ayoubi S, Nourbakhsh F. Prediction of soil enzymes activity by digital terrain analysis: comparing artificial neural network and multiple linear regression models. Environ Eng Sci. 2012. https://doi.org/10.1089/ees.2011.0313.
132. Larsen PE, Field D, Gilbert JA. Predicting bacterial community assemblages using an artificial neural network approach. Nat Methods. 2012. https://doi.org/10.1038/nmeth.1975.
133. Santos EC, Armas ED, Crowley D, Lambais MR. Artificial neural network modeling of microbial community structures in the Atlantic Forest of Brazil. Soil Biol Biochem. 2014. https://doi.org/10.1016/j.soilbio.2013.10.049.
134. Ebrahimi M, Sinegan AAS, Sarikhani MR, Mohammadi SA. Comparison of artificial neural network and multivariate regression models for prediction of Azotobacteria population in soil under different land uses. Comput Electron Agric. 2017. https://doi.org/10.1016/j.compag.2017.06.019.
135. Ramirez KS, Knight CG, De Hollande M, Brearley FQ, Constantinides B, Cotton A, Creer S, Crowther TW, Davison J, Delgado-Baquerizo M, Dorrepaal E, Elliott DR, Fox G, Griffiths RI, Hale C, Hartman K, Houlden A, Jones DL, Krab EJ, Maestre FT, McGuire KL, Monteux S, Orr CH, van der Putten WH, Roberts IS, Robinson DA, Rocca JD, Rowntree J, Schlaeppi K, Shepherd M, Singh BK, Straathof AL, Bhatnagar JM, Thion C, van der Heijden MGA, De Vries FT. Detecting macroecological patterns in bacterial communities across independent studies of global soils. Nat Microbiol. 2018. https://doi.org/10.1038/s41564-017-0062-x.
136. Iadanza E, Fabbri R, Bašić-ČiČak D, Amedei A, Telalovic JH. Gut microbiota and artificial intelligence approaches: a scoping review. Health Technol. 2020. https://doi.org/10.1007/s12553-020-00486-7.
137. Oh M, Zhang L. DeepMicro: deep representation learning for disease prediction based on microbiome data. Sci Rep. 2020. https://doi.org/10.1038/s41598-020-63159-5.
138. Zeng T, Yu X, Chen Z. Applying artificial intelligence in the microbiome for gastrointestinal diseases: a review. J Gastroenterol Hepatol. 2021. https://doi.org/10.1111/jgh.1550.
139. Kumar P, Sinha R, Shukla P. Artificial intelligence and synthetic biology approaches for human gut microbiome. Crit Rev Food Sci Nutr. 2022. https://doi.org/10.1080/10408398.2020.1850415.
140. Vilne B, Ķibilds J, Siksna I, Lazda I, Valciņa O, Krūmiņa A. Could artificial intelligence/machine learning and inclusion of diet-gut microbiome interactions improve disease risk prediction? Case study: coronary artery disease. Front Microbiol. 2022. https://doi.org/10.3389/fmicb.2022.627892.
141. Namkung J. Machine learning methods for microbiome studies. J Microbiol. 2020. https://doi.org/10.1007/s12275-020-0066-8.
142. Curry KD, Nute MG, Treangen TJ. It takes guts to learn: machine learning techniques for disease detection from the gut microbiome. Emerg Top Life Sci. 2021. https://doi.org/10.1042/ETLS20210213.
143. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, Swenson TL, Van Goethem MW, Northen TR, Vazquez-Baeza Y, Wang M, Bokulich NA, Watters A, Song SJ, Bonneau R, Dorrestein PC, Knight R. Learning representations of microbe–metabolite interactions. Nat Methods. 2019. https://doi.org/10.1038/s41592-019-0616-3.
144. Tataru CA, David MM. Decoding the language of microbiomes using word-embedding techniques, and applications in inflammatory bowel disease. PLoS Comput Biol. 2020. https://doi.org/10.1371/journal.pcbi.1007859.
145. Costello Z, Martin HG. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. NPJ Syst Biol Appl. 2018. https://doi.org/10.1038/s41540-018-0054-3.
146. Metwally AA, Yu PS, Reiman D, Dai Y, Finn PW, Perkins DL. Utilizing longitudinal microbiome taxonomic profiles to predict food allergy via long short-term memory networks. PLoS Comput Biol. 2019. https://doi.org/10.1371/journal.pcbi.1006693.
147. Liu Y, Zhu J, Wang H, Lu W, Lee YK, Zhao J, Zhang H. Machine learning framework for gut microbiome biomarkers discovery and modulation analysis in large-scale obese population. BMC Genomics. 2022. https://doi.org/10.1186/s12864-022-09087-2.
148. Li P, Luo H, Ji B, Nielsen J. Machine learning for data integration in human gut microbiome. Microb Cell Fact. 2022. https://doi.org/10.1186/s12934-022-01973-4.
149. Hernández Medina R, Kutuzova S, Nielsen KN, Johansen J, Hansen LH, Nielsen M, Rasmussen S. Machine learning and deep learning applications in microbiome research. ISME Commun. 2022. https://doi.org/10.1038/s43705-022-00182-9.
150. Ghannam RB, Techtmann SM. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Comput Struct Biotechnol J. 2021. https://doi.org/10.1016/j.csbj.2021.01.028.
151. Xia Y. Statistical normalization methods in microbiome data with application to microbiome cancer research. Gut microbes. 2023. https://doi.org/10.1080/19490976.2023.2244139.
152. Hossain MR, Timmer D. Machine learning model optimization with hyper parameter tuning approach. Glob J Comput Sci Technol D Neural Artif Intell. 2021;21(2):31.
153. Papoutsoglou G, Tarazona S, Lopes MB, Klammsteiner T, Ibrahimi E, Eckenberger J, Novielli P, Tonda A, Simeon A, Shigdel R, Béreux S, Vitali G, Tangaro S, Lahti L, Temko A, Claesson MJ, Berland M. Machine learning approaches in microbiome research: challenges and best practices. Front Microbiol. 2023. https://doi.org/10.3389/fmicb.2023.1261889.
154. Deng Z, Zhang J, Li J, Zhang X. Application of deep learning in plant–microbiota association analysis. Front Genet. 2021. https://doi.org/10.3389/fgene.2021.697090.
155. Cuomo S, De Rosa M, Piccialli F, Pompameo L, Vocca V. A numerical approach for soil microbiota growth prediction through physics-informed neural network. Appl Numer Math. 2024. https://doi.org/10.1016/j.apnum.2024.08.025.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.3 Supplementary materials

Table S1 Results of the research with combinations of terms and Boolean operator “AND”

Combinations	Google Scholar	Web of Science	Scopus
artificial intelligence AND soil microbial community	1010	0	3
artificial intelligence AND soil health prediction	26	0	0
artificial intelligence AND soil microbiome monitoring	3	0	0
artificial intelligence AND soil health assessment	269	0	1
artificial intelligence AND soil microbiome	945	6	8
artificial intelligence AND soil microbial indicators	4	0	0
soil microbial community AND soil health prediction	2	0	0
soil microbial community AND soil microbiome monitoring	6	0	0
soil microbial community AND soil health assessment	731	3	5
soil microbial community AND soil microbiome *	\	\	\
soil microbial community AND soil microbial indicators *	\	\	\
soil health prediction AND soil microbiome monitoring	0	0	0
soil health prediction AND soil health assessment	7	0	0
soil health prediction AND soil microbiome	3	0	0
soil health prediction AND soil microbial indicators	0	0	0
soil microbiome monitoring AND soil health assessment	0	0	0
soil microbiome monitoring AND soil microbiome *	\	\	\
soil microbiome monitoring AND soil microbial indicators	0	0	0
soil health assessment AND soil microbiome	365	1	3
soil health assessment AND soil microbial indicators	34	0	0
soil microbiome AND soil microbial indicators	103	0	0

(*redundant)

Table S2 Brief compendium with the main terms related to AI, named in this review, for metrics, NN architecture, statistical, and training techniques

Glossary		
Terms	What is	Use
Accuracy	A statistical metric	measures the proportion of correctly classified examples out of the total number of examples in the validation dataset.
Autoencoders	An artificial neural network	designed to learn efficient and typically lower-dimensional representations of data, often for dimensionality reduction or feature learning.
Bayesian optimization	A ML technique	used to find the optimal parameters or hyperparameters of a complex black-box function, minimizing the number of observations while converging rapidly to the optimal solution.
Convolutional Neural Networks (CNNs)	An artificial neural network	particularly effective for analyzing visual imagery, widely used in image and video recognition, image classification, medical image analysis, and other tasks involving large amounts of visual data.
Cross-validation	A statistical method	define a dataset to “test” the model in the training phase to validate the model, to limit problems like overfitting, and underfitting, and provide an insight

		into how the model will generalize to an independent dataset.
F1-score	A statistical metric	often used in binary classification tasks, where the goal is to classify instances into one of two classes, typically labeled as positive or negative.
Gradient Boosting Machine (GBM)	A ML technique	used for classification and regression tasks, builds a series of weak learners (typically decision trees) sequentially, with each new learner correcting the errors made by the previous ones.
Grid search	A ML technique	it ensures a systematic exploration of the hyperparameter space and helps in finding the optimal for a given learning algorithm and dataset.
Hyperparameter optimization	A ML technique	the process of finding the best set to optimize the performance of the machine learning model.
k-fold cross validation	A statistical method	used to assess the performance of a machine learning model and to evaluate its generalization ability, partitioning the dataset into *k* equally sized folds.
k-nearest neighbors	A ML technique	non-parametric, supervised learning method that uses proximity to make classifications or predictions about the grouping of an individual data point.
Logarithmic transformation	A mathematical operation	involves applying the logarithm function to a set of data points, commonly used to transform skewed data distributions or to stabilize variance in the data.
Minimal values	A mathematical optimization	considered when analyzing data for outlier detection, minimum variance, or in methods like Min-Max scaling, which involves scaling features to a fixed range based on the minimum and maximum values.
Mean Squared Error (MSE)	A statistical metric	used to measure the average squared difference between the observed values and the predicted values produced by a model, for evaluating the performance of regression models.
Multiple imputation	A statistical technique	used to create multiple complete datasets by filling in missing values several times, analyzing each dataset separately, and then combining the results to get final estimates.
Artificial Neural Networks (NNs)	A computational model	inspired by the human brain, consisting of layers of interconnected nodes or "neurons" that process and learn from data. Each connection has a weight that adjusts during training to minimize errors, enabling the network to make predictions or classify information.
Normalization	A statistical technique	used in data analysis and machine learning to scale numeric features to a standard range, bringing all features to a similar scale, preventing features with large values from dominating the learning process, and ensuring that the model can learn from all features equally.
Permutation importance	A ML technique	used to understand which features contribute most to the predictions of a model, evaluating how the predictions of the model change if the values of a single feature are randomly shuffled, thereby disrupting the relationship between that feature and the target outcome.
Precision	A statistical metric	defined as the correctly predicted positive observations among all the cases that the model predicted as positive, used particularly in classification tasks, to evaluate the accuracy of the positive predictions made by the model.
Principle Component Analysis (PCA)	A statistical technique	commonly used in data analysis and machine learning to simplify complex datasets with many

		variables, reducing the number of dimensions without losing critical information.
Random Forest (RF)	A ML algorithm	used for classification and regression tasks, creates a "forest" of decision trees, where each tree is trained on a random subset of the data. This randomness helps in making the model more robust and less prone to overfitting.
Random Search	A ML technique	used for optimization particularly useful when the hyperparameter space is high-dimensional or when computational resources are limited.
Rarefaction	A statistical technique	used to estimate the species richness or diversity of a population based on a standardized sample size, comparing species diversity in different samples, particularly when the samples have different sizes.
Recall	A ML technique	used in classification tasks to measure the ability of a model to identify all relevant instances in the dataset, focusing on the accuracy of the positive predictions made by the model.
Recurrent Neural Networks (RNNs)	An artificial neural network	particularly suited for processing sequences of data, such as time series, speech, text, or any other form of sequential information, characterized by loops, allowing information to persist and influencing the network's output based on previous computations.
Root Mean Squared Error (RMSE)	A statistical metric	used to evaluate the performance of a regression model, measuring the average deviation of the predicted values from the actual values.
SHAP values	A model metric	used to explain the output of machine learning models by attributing the contribution of each feature to the model's predictions.
Support Vector Machines (SVM)	A ML algorithm	are primarily used for classification tasks, but can also be adapted for regression, important for their effectiveness in high-dimensional spaces and their ability to handle both linear and nonlinear decision boundaries.
Total-sum scaling	A statistical technique	useful in preparing data for analysis by making sure that the features or samples are on a comparable scale, normalizing the data relative to the total sum, and turning absolute measurements into relative measurements, which represent proportions or percentages of the whole.
Tree-based model	A ML algorithm	uses a decision tree structure to make predictions, by making sequential, hierarchical decisions about the data features.
t-distributed stochastic neighbor embedding (t-SNE)	A ML technique	used to visualize high-dimensional data in lower-dimensional space, typically two or three dimensions, particularly useful for exploring and understanding the structure of complex datasets by preserving the local and global relationships between data points.
Uniform Manifold Approximation and Projection (UMAP)	A ML technique	preserves local structures well, but it also balances the preservation of global structures, making it useful for a broader range of data analysis tasks.
Word-embedding techniques	A ML technique	used to represent words as dense vectors of real numbers in a high-dimensional space, capturing semantic and syntactic relationships between words, allowing machines to process and understand human language more effectively.

Chapter 3 – Original research

3.1 Introduction

Building on the conceptual framework outlined in the previous chapter, this section presents the experimental core of this doctoral research. The study focuses on the characterization of soil microbial communities in an agricultural system subjected to different management practices, with the aim of understanding how these practices shape microbial composition, structure, and ecological functions over time.

Instead of relying on classical ecological metrics, the study explored the potential of AI and ML algorithms to detect hidden relationships among treatments, microbial taxa, and soil health indicators. These approaches aimed to move beyond descriptive characterization toward explorative algorithms, supporting the identification of microbial groups associated with improved soil fertility and sustainable productivity.

This study has been accepted for publication in the Special Issue "Smart Agriculture – BrIAS Edition" of *Physiologia Plantarum*, following a minor revision process, and is currently awaiting final publication. It is presented here in its accepted form, with the supplementary material at the end of the main text. The work was also presented at the II Conference on Smart Agriculture, organized by the Brussels Institute for Advanced Studies (BrIAS) and held in Brussels on 6–7 February 2025, where it contributed to the broader scientific dialogue on the role of AI in transforming agrifood systems.

Overall, this work provides a comprehensive view of how soil microbiota responds to management inputs and highlights the role of microbial communities in enabling data-driven, sustainable agriculture.

3.2 Machine Learning Approaches to Assess Soil Microbiome Dynamics and Bio-Sustainability

Roberta Pace[1,2,*], Maurilia M. Monti[2*], Salvatore Cuomo[3], Antonio Affinito[4], Michelina Ruocco[2]

[1]Department of Biology of the University of the study of Naples "Federico II" – Italy
[2]Institute of Sustainable Plant Protection of National Research Council (IPSP – CNR) – Portici, Naples, Italy
[3]Department of Mathematics and Applications "Renato Caccioppoli" of the University of the study of Naples "Federico II" – Italy
[4] EVJA s.r.l. - Via Benedetto Brin, 63, 80142 Napoli NA
[*]Corresponding authors: roberta.pace@unina.it; mauriliamaria.monti@cnr.it

Abstract

Understanding soil microbiota dynamics is essential for enhancing bio-sustainability in agriculture, yet the complexity of microbial communities hampers prediction of their functional roles. Artificial Intelligence (AI) and Machine Learning (ML) offer powerful tools to analyse high-dimensional microbiome data generated by high-throughput sequencing.

Here, we apply unsupervised AI-based algorithms to uncover microbial patterns that are not immediately recognisable but are crucial for characterising the biological status of agricultural soils. Soil samples were collected from a site in Northern Italy managed under four strategies: conventional farming without organic matter (C), with organic matter (C+O), with beneficial microorganisms but without organic matter (M), and with both beneficial microorganisms and organic matter (M+O). Metagenomic amplicon sequencing of the 16S ribosomal RNA (rRNA) gene and the internal transcribed spacer (ITS) region was used to profile bacterial and fungal communities.

Principal Component Analysis (PCA), k-means clustering and t-distributed Stochastic Neighbour Embedding (t-SNE) revealed coherent temporal trajectories in both datasets, with sampling time and crop presence emerging as dominant drivers of community assembly and only subtle compositional shifts attributable to treatments. Fungal communities exhibited higher plasticity and a stronger response to management than bacterial communities, which converged towards a stable oligotrophic core.

Our findings highlight the complementary roles of fungal and bacterial guilds and show that unsupervised ML-based workflows provide an effective framework to disentangle temporal and treatment effects in complex microbiome datasets. This exploratory study lays the groundwork for future predictive models aimed at identifying microbial indicators of soil biological status and supporting bio-sustainable agronomic decisions.

1.Introduction

Soil microbiota encompasses a diverse community of microorganisms inhabiting the soil, including bacteria, fungi, viruses, and algae. These communities are integral components of the soil ecosystem, present across all soil types, with their abundance, diversity, and composition significantly variable.

Microorganisms adjust their community structure and functions in response to specific environmental and physicochemical soil conditions and are present across various ecosystems, including agroecosystems, forests, grasslands, and even deserts (He et al. 2023; Cole et al. 2024; Zong et al. 2024; Burrill et al. 2025). Soil microbiota is distributed throughout the soil profile, with its highest concentrations typically found in the rhizosphere (Omotayo and Babalola 2021), thanks to the roots releasing root exudates, representing a readily available source of carbon and energy for microorganisms. This enrichment creates a biological 'hotspot': microbes proliferate more rapidly and establish highly dense and functionally distinct communities (Kuzyakov and Razavi 2019). They are also present in bulk soil, within aggregates, and in soil pores. Notably, a single gram of soil can contain up to 10 billion microorganisms spanning thousands of distinct species (Pandey et al. 2024).

Soil should be regarded as a non-renewable resource and a vital and dynamic ecosystem in which organisms and microorganisms interact in a bidirectional manner (Nannipieri 2020; Pedrinho et al. 2024). These interactions are particularly crucial, as they profoundly influence the multifunctionality of the soil ecosystem (Wang et al. 2024; Hossain et al. 2025). A healthy soil sustains essential ecosystem services, including water retention, nutrient cycling, carbon sequestration, and erosion control. Moreover, it plays a critical role in food production by providing nutrients, water, and structural support necessary for plant growth (Kopittke et al. 2024). Given these considerations, a fundamental question emerges: how can soil health be effectively assessed? Assessing soil health successfully involves evaluating its biological, chemical, and physical properties to determine how well it supports plants, animals, and humans. Physical indicators, in particular soil texture, aggregate stability, porosity, and bulk density give information on structure, compaction and water dynamics, that influence root growth, water retention, and gas exchange. Chemical indicators, such as pH, cation exchange capacity, organic matter content, nutrient levels (N, P, K, micronutrients) and presence of contaminants, and electrical conductivity, reflect fertility, nutrient cycling, and pollution status. But, as largely discussed in scientific literature, an effective assessment of soil health must encompass a detailed analysis of its biological components, with particular emphasis on the soil microbial community, which also plays a pivotal role in mediating the soil–plant–ecosystem relationship (Doran and Zeiss 2000; Ferris and Tuomisto 2015; Lehman et al. 2015; Fierer et al. 2021; Bhaduri et al. 2022). The structure and functionality of these microbial assemblages critically influence nutrient cycling by decomposing organic material, thereby releasing plant-available nutrients and contributing to the formation of humus, which enhances soil water retention (Lehmann and Kleber 2015). They also mediate key biochemical processes such as oxidation, reduction, solubilization, and chelation, facilitating nutrient availability (Martinez et al. 2013). By producing extracellular substances, these microorganisms contribute to soil aggregation, strengthening soil structure and stability (Hartmann and Six 2023; Mueller et al. 2024). Additionally, soil microbiota establishes interactions with plants that promote growth and improve resistance to biotic and abiotic stresses (Hou et al. 2021; Omae and Tsuda 2022; Kapoor et al. 2024).

Advancements in molecular techniques, particularly metagenomics, have revolutionized the study of soil microbiota by enabling a comprehensive assessment of microbial communities. This approach

allows the analysis of genetic material directly from soil samples, providing insights into microbial diversity and abundance (Taraboletti et al. 2023), without the need for microbiological culture and thus overcoming the limitations posed by unculturable microorganisms. By facilitating comparisons across different environmental conditions and agricultural treatments, metagenomics serves as a powerful tool for detecting shifts in microbial populations that directly impact soil health and ecosystem functionality (Masuda et al. 2024).

This study focuses on soil microbiota, particularly fungal and bacterial communities within an intensively managed greenhouse system cultivating *Valerianella locusta* (lamb's lettuce). In such baby-leaf systems, short crop cycles, continuous re-sowing and limited crop rotation create a highly dynamic environment, in which temporal succession and rhizosphere effects may overshadow the impact of agronomic treatments. Yet, the relative contribution of the sampling time, crop presence and management to soil microbial dynamics remains poorly understood in these contexts.

Here, we investigate soil microbial communities through DNA-based profiling and advanced computational analyses. We use AI-based exploratory tools, Principal Component Analysis (PCA), K-means clustering, and t-distributed Stochastic Neighbour Embedding (t-SNE), to identify major patterns and discrete microbial "states" across treatments and sampling times. This unsupervised ML framework is designed to disentangle temporal vs treatment effects, compare the responses of fungal and bacterial communities, and provide a structured basis for future predictive modelling.

The ultimate goal is to pave the way for a predictive model that leverages Artificial Intelligence (AI) approaches, including Machine Learning (ML) and Deep Learning (DL) algorithms, to identify microbial patterns contributing to soil fertility. The analysis of soil microbiome data through artificial intelligence (AI) involves several critical steps (Pace et al. 2025):

1. **Data Collection** – Ensuring the accuracy and representativeness of soil samples.
2. **Pre-processing and Transformation** – Cleaning, standardizing, and preparing data for computational analysis.
3. **AI Model Selection** – Identifying appropriate algorithms for predictive modelling.
4. **Training and Validation** – Developing and refining the model using sample data to enhance reliability.
5. **Model Performance Evaluation** – Assessing the model's accuracy and applicability in predicting microbial community dynamics.

Following this framework, we structured our research into three main phases: data collection, data analysis (pre-processing and transformation), and model implementation. In the present work, we report on the outcomes of the first two phases. Specifically, we address the following questions: (i) how do soil fungal and bacterial communities evolve over successive crop cycles of *V. locusta*? (ii) to what extent do the four agronomic treatments, including microbial inoculation and organic matter addition, alter these temporal trajectories? and (iii) do fungal and bacterial communities differ in response to the same management regime? By answering these questions, we aim to clarify the ecological dynamics of soil

microbiota in greenhouse baby-leaf production and to illustrate how unsupervised ML approaches can support soil health assessment under agronomic conditions.

2.Materials and Methods

2.1 Site description and treatments

Soil sampling and environmental data collection were conducted from January to September 2023 in a commercial farm belonging to the Producer Organization (OP) "Sole&Rugiada", located in Bagnolo Mella (Brescia, Lombardy, Italy). The farm specializes in greenhouse cultivation of baby leaf crops, and the study was carried out on *Valerianella locusta*.

According to the experimental design, four treatments were established: (i) conventional farming without organic matter (C), (ii) conventional farming with the addition of organic matter (C+O), (iii) conventional farming with the addition of beneficial microorganisms but without organic matter (M), and (iv) conventional farming with the addition of both beneficial microorganisms and organic matter (M+O). The term "addition of beneficial microorganisms" refers to the application, via irrigation, of the commercial product "RyzoPepUp" by Samagri S.r.l. (Cava de' Tirreni, NA Italy), which contains rhizosphere bacteria and *Trichoderma* spp. This formulation is designed to promote the development of both the root system and the aerial parts of the crop. The product is delivered in an organic matrix consisting of a non-composted plant-based soil improver. In this experiment, RyzoPepUp was applied at a commercial dose and timing recommended by the manufacturer. Consequently, the presence or absence of the microbial inoculant (treatments M and M+O vs. C and C+O) was treated as a categorical management factor, rather than as a graded continuous input signal. Treatments with microorganisms were applied throughout the sampling period, with one application per crop cycle.

2.2 Soil sampling scheme and experimental design

Two baseline samples of bare soil were collected, one from the plots designated for the microbial inoculum (M/M+O) and one from the control plots (C/C+O). At that stage the plots had only been demarcated and no treatments had yet been implemented. Subsequent soil samplings were made at the end of each crop cycle. Sampling time points are thus denoted t0, t1, t2, t3, t4, and t5, with t0 representing the shared baseline and t1 to t5 the subsequent samplings.

For each treatment and time point, sampling points were arranged according to a standardized grid consisting of two diagonals across each plot, with five evenly spaced points per diagonal (Figure 1). Each diagonal yielded one composite soil sample, resulting in two samples per plot. From these two field replicates collected per treatment, a third composite sample was prepared by homogenization to obtain a more representative biological replicate. As a result, three biological replicates were obtained for each treatment at each time point. All soil samples were collected in clean plastic bags, stored at 4 °C, and processed within one week from collection. The sampling depth corresponded to the soil layer occupied by the root system, which, in the case of baby leaf crops, does not exceed 30 cm.

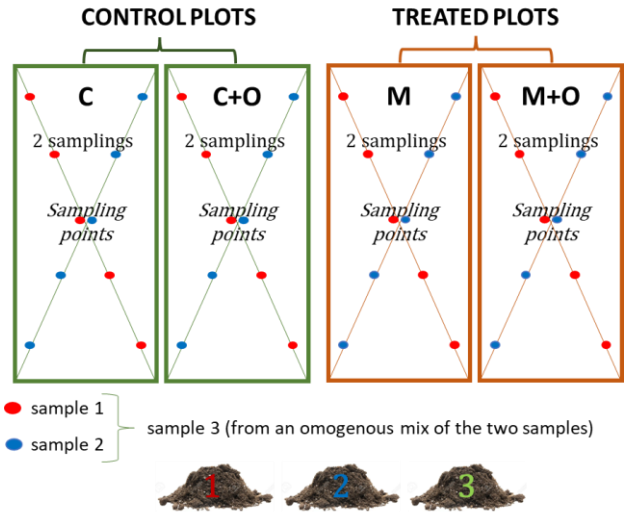


Figure 1. Schematic representation of the soil-sampling design in the greenhouse experiment. In each of the four plots, two composite soil samples were collected by pooling cores at fixed positions along two diagonal transects (red and blue dots). A third soil sample was then obtained by mixing equal amounts (1:1) of the "blue" and "red" composites, yielding three samples (1–3) per plot.

2.3 DNA extraction and amplicon preparation

The biological characteristics of the soil samples were assessed through metagenomic analysis. Genomic DNA of bacteria and fungi was extracted from each soil sample and subsequently amplified in triplicate. DNA extraction was carried out using the NucleoSpin® Microbial DNA kit (Macherey-Nagel, Germany), following the manufacturer's protocol for genomic DNA isolation from microbial communities. For bacterial community profiling, the universal primers 27F (forward: 5'-AGAGTTTGATCCTGGCTCAG-3') and 1492R (reverse: 5'-GCTTACCTTGTTACGACTT-3') (Frank et al. 2008) targeting the 16S rRNA gene were used. For fungal community analysis, the universal primers ITS1 (forward: 5'-TCCGTAGCTGAACCTGCGG-3') and ITS4 (reverse: 5'-TCCTCCGCTTATTGATATGC-3') (Mirhendi et al. 2007) targeting the ITS region were employed.

PCR amplification was carried out in a 50 µL reaction volume containing: 10 µL of 5X Colorless GoTaq® Reaction Buffer with $MgCl_2$ (Promega, USA), 1 µL of forward primer (10 µM), 1 µL of reverse primer (10 µM), 1 µL of dNTP mix (10 mM each; Promega, USA), 0.25 µL of GoTaq® DNA Polymerase (5 U/µL; Promega, USA), 0.5 µL of DNA template (10 ng/µL), and 36.25 µL of sterile ultrapure water. Thermal cycling was conducted using a Mastercycler® Nexus X2 (Eppendorf®, Germany) with the following conditions:

– For ITS amplification: initial denaturation at 95 °C for 5 min; 35 cycles of 95 °C for 30 s, 58 °C for 30 s, 72 °C for 50 s; final extension at 72 °C for 10 min.

– For 16S amplification: initial denaturation at 95 °C for 5 min; 35 cycles of 95 °C for 30 s, 48 °C for 30 s, 72 °C for 90 s; final extension at 72 °C for 10 min.

PCR products were visualized on a 1% agarose gel prepared in TAE 1% buffer and stained with SYBR Safe DNA Gel Stain (10%; Invitrogen, Thermo Fisher Scientific, USA). A 100 bp DNA ladder (New England Biolabs Inc., UK) was used as a molecular size marker. Following electrophoresis, the amplicons were purified using the NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel, Germany), according to the manufacturer's protocol. The concentration of purified PCR products (ng/μL) was measured using both the NanoDrop™ One spectrophotometer (Thermo Fisher Scientific, USA) and the Qubit™ Fluorometer (Invitrogen, Thermo Fisher Scientific, USA) for improved quantification accuracy.

2.4 Oxford Nanopore sequencing

Amplicon sequencing was performed by using Oxford Nanopore Technologies (ONT), which allows the analysis of near full-length marker genes, improving taxonomic resolution compared to short-read platforms. The three PCR replicates from each soil sample were pooled. Library preparation was performed using the Rapid Barcoding Kit 24 (SQK-RBK114.24; Oxford Nanopore Technologies, UK), and sequencing was carried out on a Flongle flow cell using the MinION platform, following the manufacturer's instructions.

2.5 Bioinformatic workflow

Raw signals were processed with Guppy, the ONT software for basecalling and demultiplexing, retaining reads with a quality score ≥Q7.

Subsequent bioinformatic processing was carried out using the EPI2ME platform (Oxford Nanopore Technologies). The workflow wf-16S, that classified 16S/18S/ITS amplicons, was applied. This workflow includes read quality control, alignment against reference databases, and taxonomic classification. Relative abundances were calculated from the classified reads. Taxonomic assignments, performed via alignment against reference databases (e.g., GenBank, NCBI), were reported primarily at the genus level, with species-level resolution considered when supported by sufficient read depth and classification confidence. The resulting abundance tables, reporting the number of reads per taxon per sample barcode, were exported in Excel format for downstream analyses.

2.6 Machine learning analytical pipeline

Preliminary data cleaning and exploratory analysis were performed using Python (v3.11) as described in Supplementary (S1), with a pipeline that integrates several unsupervised machine learning techniques, namely Principal Component Analysis (PCA), K-means clustering and t-distributed Stochastic Neighbour Embedding (t-SNE), to explore patterns and structure within complex microbial datasets and identify natural groupings among samples and potential microbial signatures associated with specific treatments or time points. For each dataset (ITS and 16S), we constructed a taxon (rows) × sample (columns) matrix of relative abundances, where each column corresponds to one treatment × time combination. These relative-abundance matrices constitute the input to the ML pipeline, while treatment and sampling time were used only as metadata to colour and annotate ordinations and were not included

as predictor variables. For the multivariate and machine-learning analyses, the relative abundances from the three biological replicates were averaged so that each treatment × time combination was represented by a single composite profile. We chose this approach to reduce noise and focus on the main patterns associated with treatments and sampling times, while still accounting for within-treatment variability at the level of data preprocessing. Metadata tables (one for ITS and one for 16S) have been included in Supplementary Materials (Table S4 and Table S5).

Principal Component Analysis (PCA) was applied to the centred and scaled relative abundance matrices to reduce dimensionality and summarise major gradients in community composition.

K-means clustering was performed on the PCA scores to identify groups of samples with similar community structure. We explored values of k from 2 to 7 and evaluated clustering performance using the average silhouette score and inspection of the elbow in the within-cluster sum of squares. For both ITS and 16S data, k = 3 provided the highest or near-highest silhouette scores while avoiding very small, unstable clusters. Moreover, the three-cluster solution yielded ecologically interpretable groups corresponding to early, intermediate and late successional stages, which were consistent across PCA and t-SNE representations. Therefore, k = 3 was retained for all subsequent analyses.

To characterise the taxa associated with each of the three K-means clusters, we combined the cluster assignments with the taxon relative-abundance tables. For each dataset (ITS and 16S), we calculated the mean relative abundance of each taxon within each cluster and used these values to identify the taxa that were most abundant and characteristic of a given cluster. This procedure is descriptive and is intended to highlight taxa that typify each microbial "state" identified by the clustering, rather than to provide a formal statistical discriminant analysis.

To visualise non-linear structure and local similarities in community composition, we applied t-distributed Stochastic Neighbour Embedding (t-SNE) to the same PCA scores used for K-means. This approach preserves neighbourhood relationships and improves the separation of clusters in a two-dimensional space. t-SNE maps were coloured by treatment, sampling time and K-means cluster assignment to facilitate the joint interpretation of temporal dynamics and management effects.

These ML algorithms improved the interpretation of ecological dynamics and treatment effects, offering a data-driven foundation for hypothesis generation. In particular, the k-means algorithm indicates the number of clusters to be generated, determined through silhouette analysis, and is based on the average (centroid) of the points belonging to each cluster (Yuan and Yang 2019); while t-SNE, is an algorithm that refers to the use of the Student's t distribution in the low-dimensional space to model similarities between points, and denotes the embedding method that aims to preserve neighbourhood relationships between data points when projecting them into a lower-dimensional space, in a probabilistic (stochastic) manner (Jung et al. 2024).

3.Results

3.1 ITS community dynamics

3.1.1 Multivariate structure and patterns

The PCA of ITS sequencing data (PC1 = 21.7%, PC2 =19.7%; cumulative variance ~41.4%) revealed a clear temporal separation of samples (Figure 2A). Baseline samples (t0) were clearly distinct from all subsequent time points, indicating a fungal community characteristic of bare soil before crop establishment. Samples collected from t1 to t4 occupied an intermediate region of the ordination space, whereas late-stage samples (t5) were slightly shifted, suggesting further restructuring of the community towards the end of the cropping cycle. Treatment-related differences were present but modest compared with the dominant temporal gradient.

K-means clustering of PCA scores partitioned the ITS dataset into three clusters (Figure 2B). Cluster 3 contained all t0 samples and represented the initial community state. Cluster 2 encompassed the majority of samples from t1 to t4, while Cluster 1 comprised mainly late-stage samples (t5) and a subset of mid-cycle samples (t2) from C and C+O treatments.

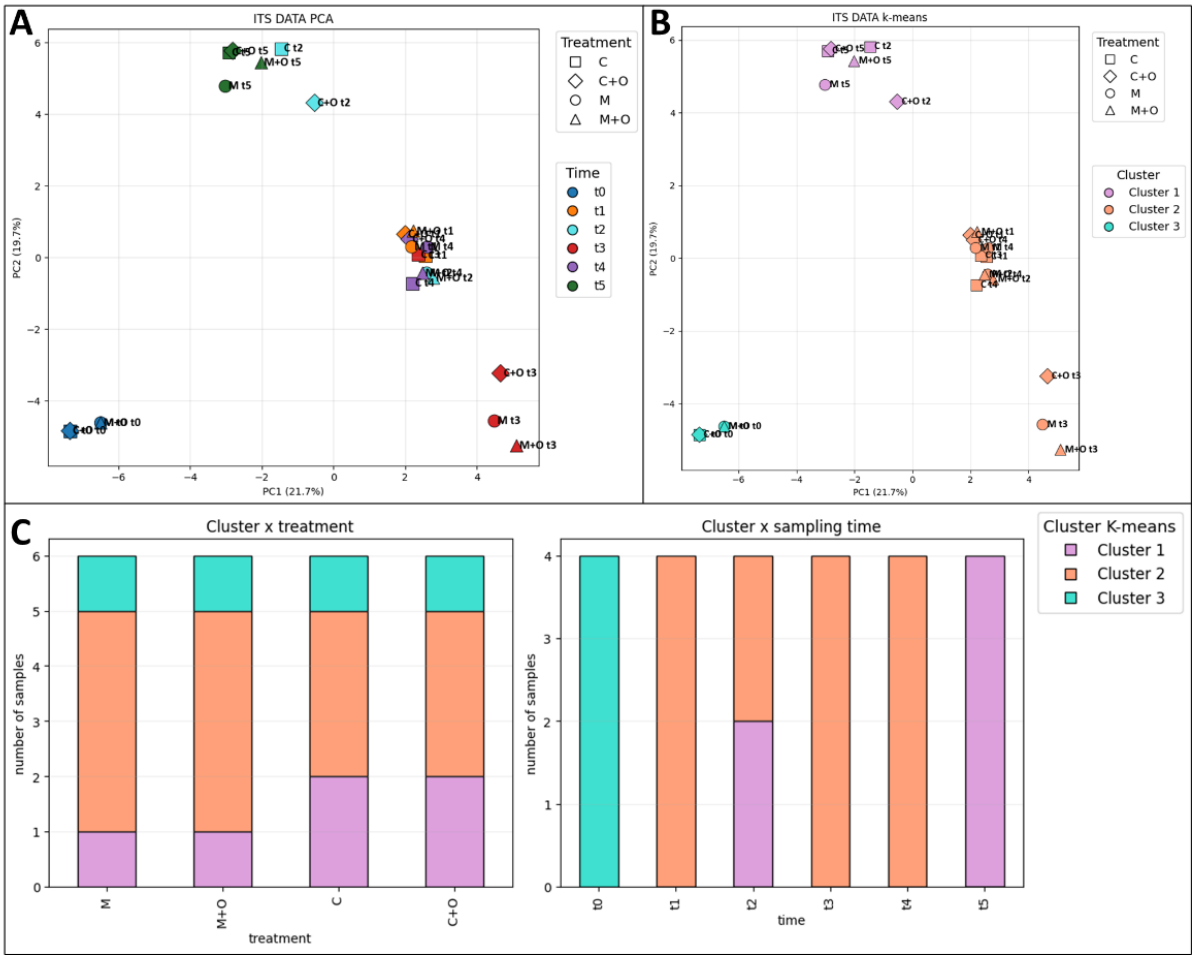


Figure 2. Multivariate structure and patterns of the ITS (fungal) community dataset. Principal Component Analysis (PCA) of the ITS (fungal) community dataset. (A) Ordination of samples along PC1 (21.7% of the variance) and PC2 (19.7%), with points coloured by sampling time (t0-blue, t1-orange, t2-cyan, t3-red, t4-violet, t5-dark green) and shaped by agronomic treatment (● M, ■ C, ▲ M+O, ◆ C+O); each point represents the mean community profile for one treatment × time combination (average of three biological replicates). (B) Same PCA ordination showing the three K-means clusters (Cluster 1 = pink, Cluster 2 = salmon, Cluster 3 = turquoise) by colour, while point shapes again denote treatments. (C) Distribution of K-means clusters in the ITS dataset, expressed as number of samples per cluster by treatment (left) and by sampling time (right).

Clustering analysis across all treatments (Figure 2C, left panel) reveals a strong predominance of Cluster 2 (salmon). The remaining clusters, Cluster 1 (pink) and Cluster 3 (turquoise), are less represented. Even if variability among treatments appears limited, and all three clusters occur across the four treatments, the proportions of the clusters seem to vary depending on the presence or absence of "RyzoPepUp". Specifically, Cluster 2 dominates in M and M+O, whereas Cluster 1 gains greater importance in C and C+O.

In Figure 2C, right panel, clusters are shown in relation to sampling time. At t0, all samples fall within Cluster 3, suggesting an initial microbial community distinct from the subsequent phases. At t1, t3, and t4, all samples group into Cluster 2, which therefore dominates the intermediate stages of the cycle. At t2, samples are evenly divided between Cluster 1 and Cluster 2, indicating a transitional stage or increased community heterogeneity at this point, reflecting a new condition in specific samples. At the

final stage (t5), all samples fall within Cluster 1, highlighting a marked reshaping of community composition compared to earlier phases.

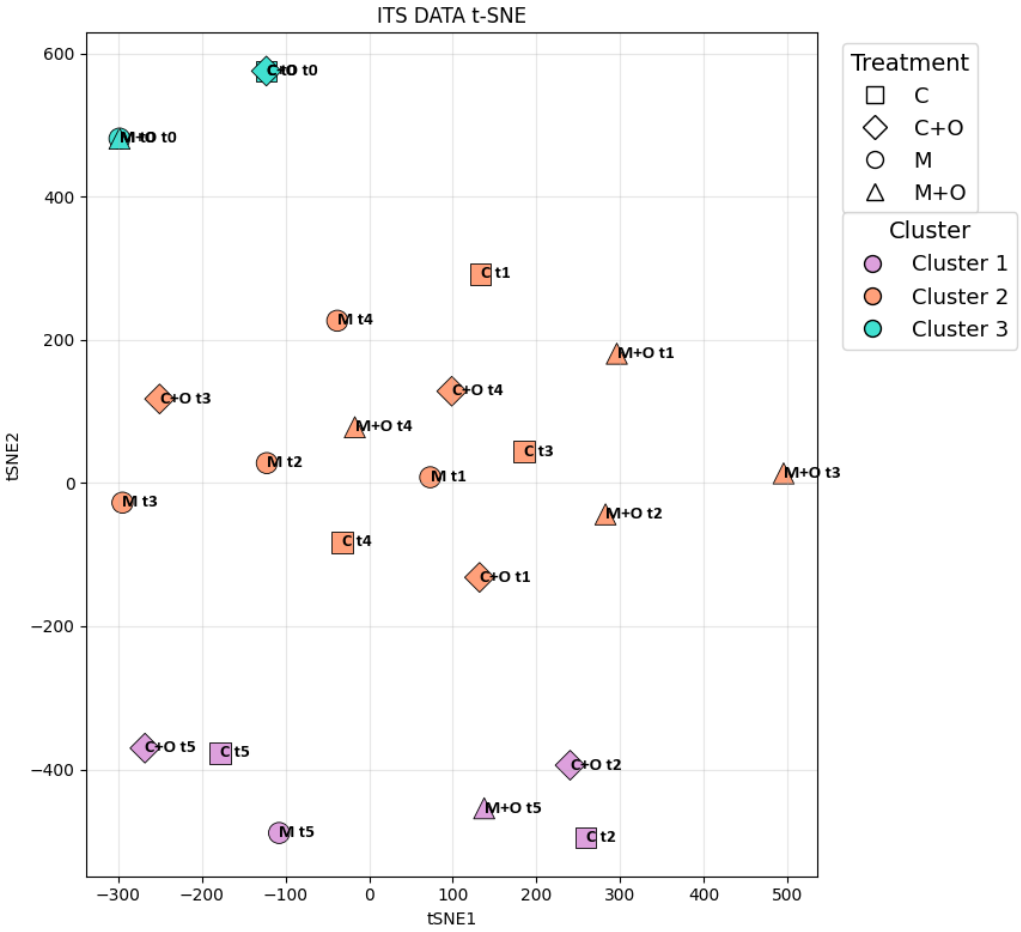


Figure 3. Non-linear ordination of the ITS (fungal) community dataset. t-distributed Stochastic Neighbor Embedding (t-SNE) plot of ITS data showing the distribution of samples in the two-dimensional t-SNE space (tSNE1 vs tSNE2). Each point represents a sample, with shapes indicating agronomic treatments (■ C, ◆ C+O, ● M, ▲ M+O) and colours corresponding to the three K-means clusters previously identified (Cluster 1 = purple, Cluster 2 = orange, Cluster 3 = turquoise).

The t-SNE analysis (Figure 3) preserved local similarities, making samples with similar features appear close together, thus revealing better cluster groupings. This analysis confirmed that sampling time, rather than treatment, was the main driver of fungal community composition. Early samples (t0) clustered separately, confirming the uniqueness of initial colonizers. Intermediate stages (t1–t4) grouped predominantly in Cluster 2, indicating convergence of communities across treatments during the central phases of the experiment. At the final stage (t5), samples formed a distinct cluster, suggesting a shift towards treatment-specific or late-successional taxa. Only a subset of t2 samples grouped in Cluster 1 together with a sample of t5, highlighting a transient divergence. Overall, t-SNE corroborated that temporal dynamics override treatment effects in shaping ITS-based community structure.

3.1.2 Cluster-level taxonomic signatures

The list of taxa that were most abundant within each fungal cluster, used to characterise their taxonomic signatures, is reported in Tables S2 (Supplementary Material S2). Below, we summarise the main taxa and functional groups that typify each cluster:

- Cluster 1 (C, C+O at t2 and M at t5) displayed a more heterogeneous composition. *Trichoderma* spp. co-occurred with putative plant-associated and pathogenic taxa (e.g. *Epicoccum* spp., *Paramacroventuria* spp.) and other functionally diverse fungi. This pattern suggests a late-successional community in which beneficial and potentially harmful taxa interact under conditions of intensified root activity, resource competition and repeated cropping.
- Cluster 2 (which includes most samples from time points t1–t4) was characterised by a strong dominance of Trichoderma spp. together with fast-growing saprophytes such as *Mortierella* spp. and *Linnemannia amoeboidea*. This assemblage likely represents an intermediate stage in which antagonistic fungi and opportunistic decomposers coexist under active rhizosphere influence and management interventions, while a background of saprotrophic fungi remains.
- Cluster 3 was clearly distinct (all samples at t0) and dominated by saprotrophic fungi typically associated with bulk-soil organic matter turnover, including *Mortierella* spp., other Mucoromycota and diverse decomposers. This profile is consistent with a decomposer-driven community in a soil rich in plant residues and undergoing active carbon turnover and nutrient mineralisation prior to crop establishment.

These differences highlight shifts from an initial saprotrophic phase (Cluster 3, dominated by *Mortierella alpina*) to communities enriched in antagonistic taxa (Clusters 1, 2), likely reflecting contrasting soil conditions and crop-associated dynamics and combining *Trichoderma* spp. with plant-associated taxa. These results demonstrate that time is the dominant factor shaping ITS community structure, while treatments exert subtler but detectable effects.

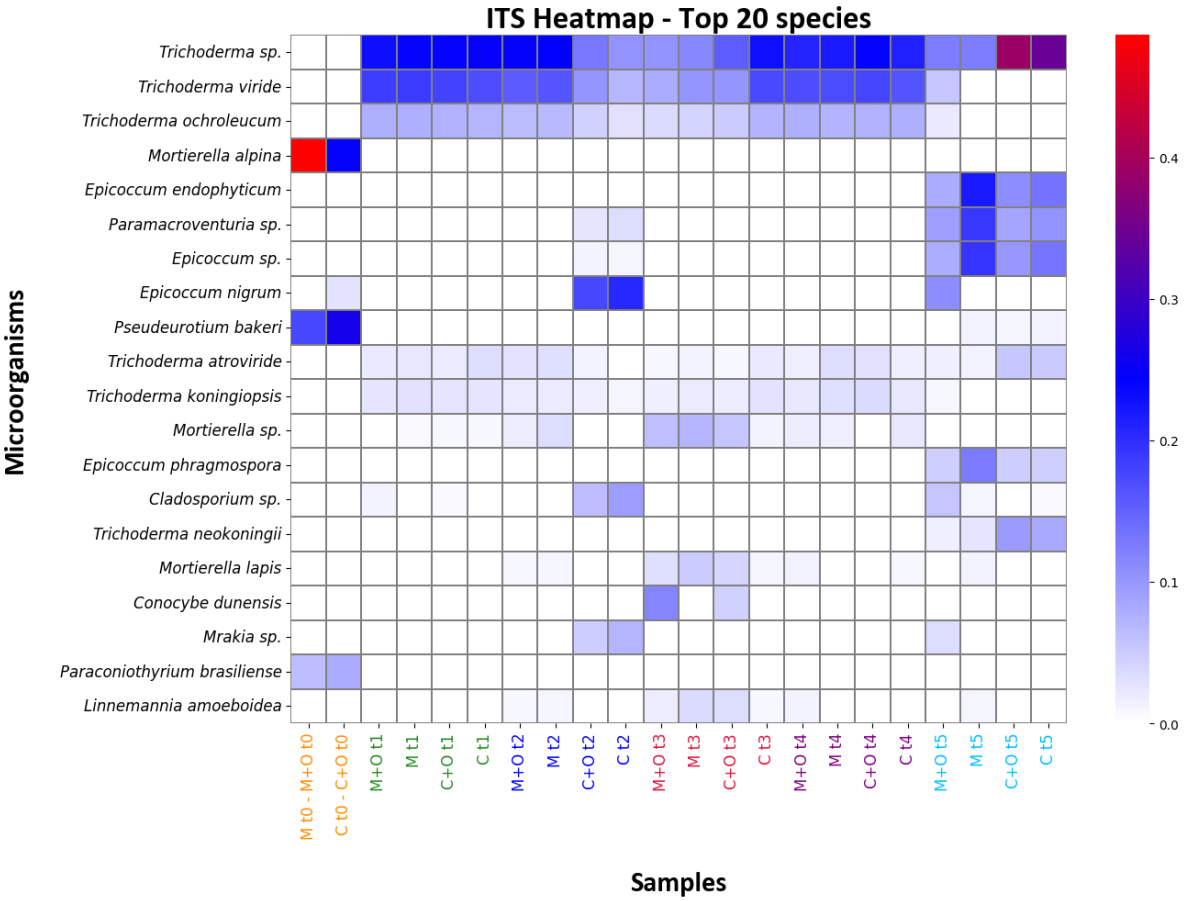


Figure 4. Heatmap of the 20 most abundant fungal species in the ITS dataset. Rows represent fungal species, while columns correspond to sample groups (treatments M, M+O, C and C+O) arranged along the temporal sequence of samplings (t0–t5); sample labels are colour-coded by sampling time. Cell colours indicate relative abundance, ranging from white (absence or very low abundance) through blue (intermediate values) to red (highest abundance).

A heatmap of the top 20 fungal species across all samples (Figure 4) provided a concise overview of compositional patterns. The heatmap illustrates the relative abundance of microorganisms across samples and reveals clear differences in the distribution of fungal taxa among treatments and sampling times. Several taxa were detected at low abundance and appeared sporadically across samples, whereas others were consistently present and more abundant, shaping distinct community patterns. Rows correspond to fungal microorganisms identified through ITS sequencing, while columns represent sequenced samples, organized by treatment and time. The side bar indicates relative abundance intensity, ranging from 0 (white) to increasing values in blue and up to the maximum in red. Notably, the t0 samples stand apart from subsequent ones, the samples collected between t1 and t4 display similar patterns, and those at t5 diverge again from the earlier time points, confirming what emerged from previous data analysis. Complete profiles (relative abundance of each taxon >0.05) and details on taxa are reported in Supplementary material (Table S4).

3.2 Bacterial (16S) community dynamics

3.2.1 Multivariate structure and patterns

The PCA based on 16S data (PC1 = 26.8%, PC2 = 20.3%, cumulative variance ~47.1%), also revealed a marked effect of time on community composition (Figure 5A). Baseline samples (t0) formed two

distinct groups reflecting the two baseline positions (C/C+O vs M/M+O areas), both separated from subsequent stages. Samples from t2 to t5 clustered more closely together, indicating a convergence of bacterial communities as the cropping cycle progressed, whereas t1 samples occupied an intermediate but slightly more dispersed position.

K-means clustering of the PCA scores identified three clusters in the bacterial dataset (Figure 5B). Cluster 3 grouped the t0 samples from C and C+O, forming the most distinctive assemblage. Cluster 2 included the t0 samples from M and M+O together with almost all samples from t2 to t5 across treatments, representing a stable core community. Cluster 1 comprised primarily t1 samples, reflecting a transient, early-stage configuration.

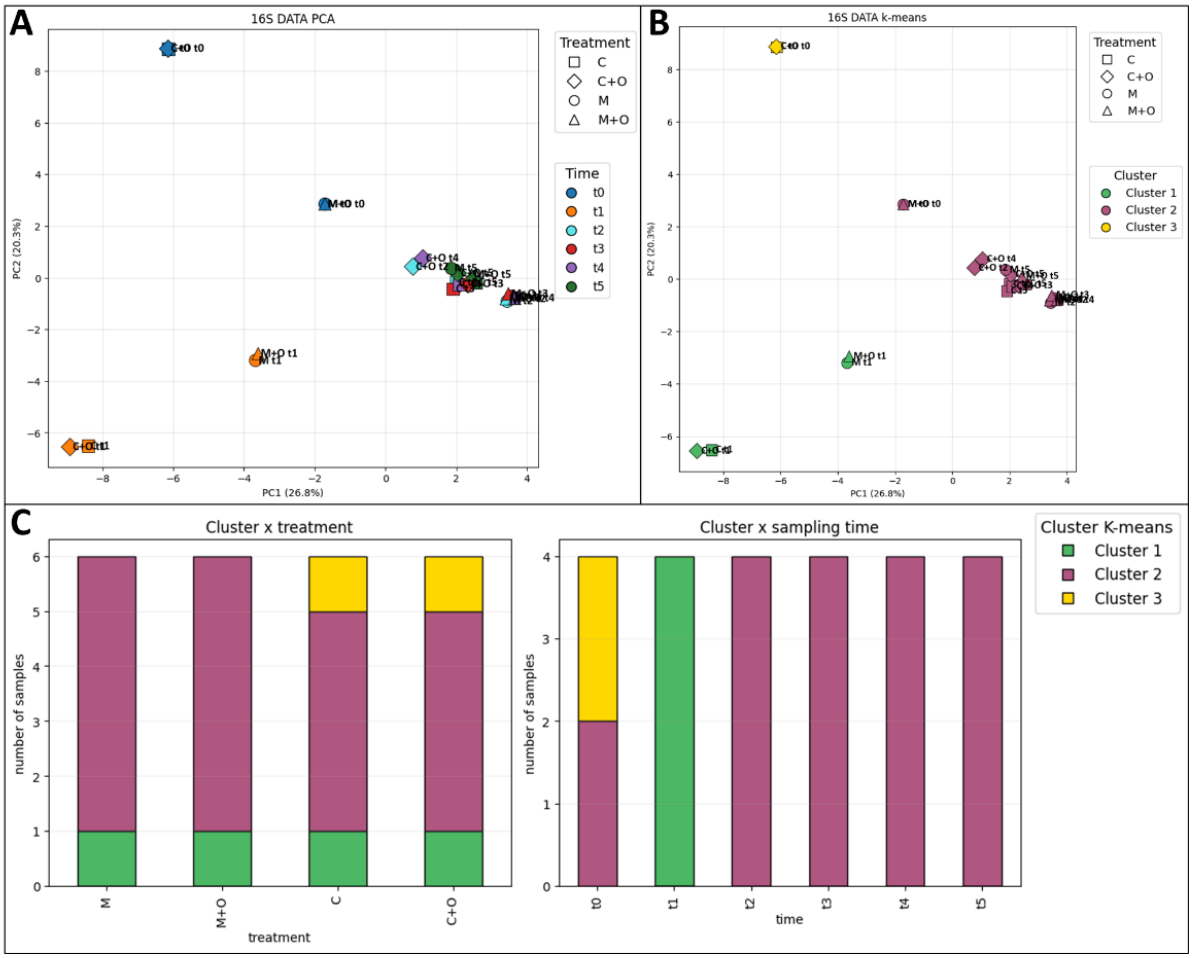


Figure 5. Multivariate structure and patterns of the 16S (bacterial) community dataset. Principal Component Analysis (PCA) of the 16S dataset. (A) Ordination of samples along PC1 (26.8% of the variance) and PC2 (20.3%), with points coloured by sampling time (t0–blue, t1–orange, t2–cyan, t3–red, t4–violet, t5–dark green) and shaped by agronomic treatment (● M, ■ C, ▲ M+O, ◆ C+O); each point represents the mean community profile for one treatment × time combination (average of three biological replicates). (B) Same PCA ordination showing the three K-means clusters (Cluster 1 = green, Cluster 2 = purple, Cluster 3 = gold) by colour, while point shapes again denote treatments. (C) Distribution of K-means clusters in the 16S dataset, expressed as number of samples per cluster by treatment (left) and by sampling time (right).

Cluster 1 and Cluster 2 are present in all the treatments (Figure 5C, left panel), with Cluster 2 (purple) largely predominant. Cluster 3 (yellow) is only present in C and C+O treatments. This indicates that microbial composition does not markedly differ among treatments. A clear temporal pattern emerges

(Figure 5C, right panel): at the initial time t0, samples are split between Cluster 2 (purple) and Cluster 3 (yellow), at t1, all samples fall exclusively into Cluster 1 (green), while from t2 to t5, all samples belong to Cluster 2 (purple).

This dynamic suggests that microbial variability is mainly driven by the temporal factor rather than by the treatment, with an initially heterogeneous community that progressively converges into a more uniform and stable profile as the cycle advances.

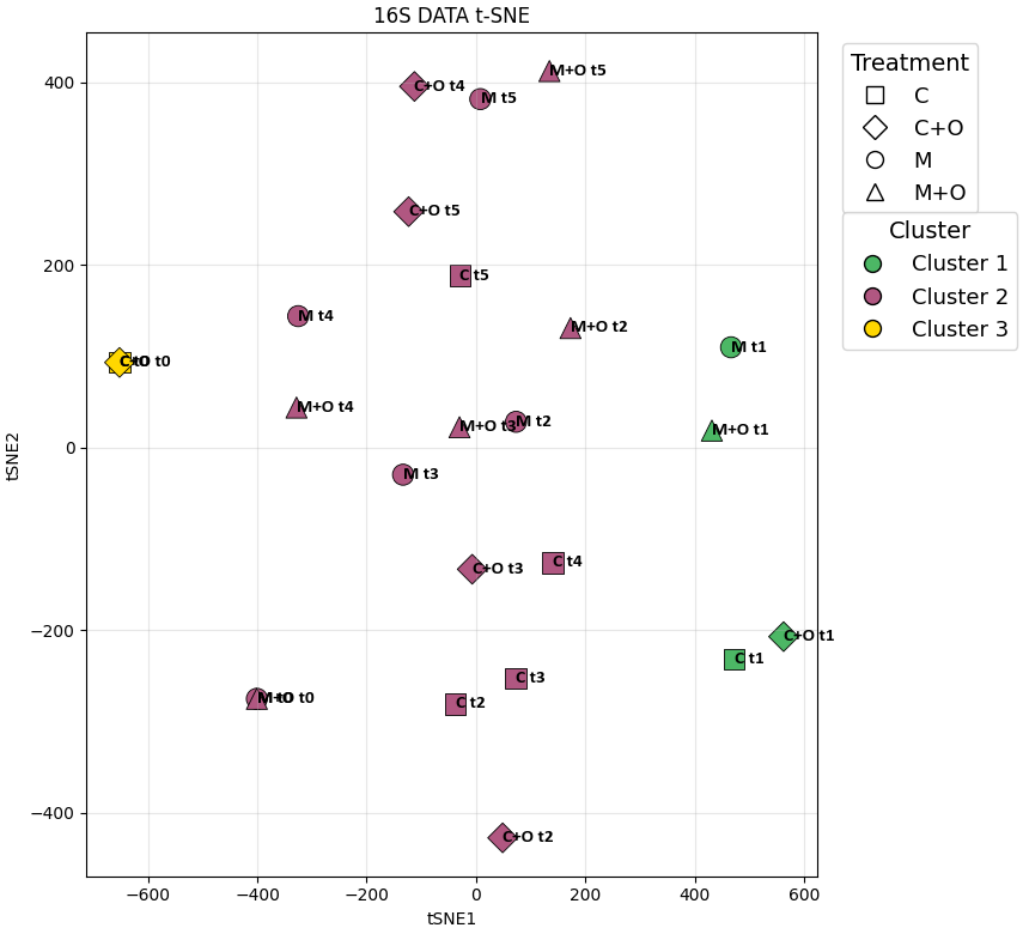


Figure 6. Non-linear ordination of the 16S (bacterial) community dataset. t-distributed Stochastic Neighbor Embedding (t-SNE) plot of 16S data showing the distribution of samples in the two-dimensional t-SNE space (tSNE1 vs tSNE2). Each point represents a sample, with shapes indicating agronomic treatments (■ C, ◆ C+O, ● M, ▲ M+O) and colours corresponding to the three K-means clusters identified (Cluster 1 = green, Cluster 2 = purple, Cluster 3 = gold).

The t-SNE plot of 16S data (Figure 6) provided a complementary, non-linear view. Baseline communities at t0 appeared as two clearly separated groups, t1 samples formed a small intermediate neighbourhood, and most samples from t2 to t5 converged in a central region, largely irrespective of treatment. This spatial arrangement confirmed that temporal dynamics were the predominant factor structuring bacterial communities, with treatments exerting only limited influence within this overall trajectory.

3.2.2 Cluster-level taxonomic signatures

Analogously, the taxa that typified each bacterial cluster, based on their mean relative abundance within clusters, are summarised in Tables S3 (Supplementary Material S3). Cluster-level taxonomic profiles showed that, in contrast to fungi, bacterial communities were dominated by a relatively stable oligotrophic core:

- Cluster 1 (mainly t1 samples) shared this core but was enriched in fast-growing, disturbance-tolerant and copiotrophic taxa, including some Bacillales and other stress-tolerant lineages. This profile is consistent with a transient colonisation phase following crop establishment and early management operations, during which opportunistic bacteria temporarily increase in abundance before the community stabilises.
- Cluster 2, which included M/M+O t0 samples and almost all samples from t2 to t5, was structured around Acidobacteriota, Actinobacteriota and other taxa typical of low-input, well-aerated soils, together with diverse Proteobacteria (e.g. *Bradyrhizobium*, *Rhizobiales*-affiliated taxa). This cluster reflects a community adapted to moderate resource availability and repeated cropping, and likely underpins functional stability in the system.
- Cluster 3, comprising C and C+O t0 samples, was the most distinctive bacterial assemblage. It included higher representation of taxa associated with wetter or more reduced microenvironments (e.g. denitrifiers and sulfur-oxidising bacteria), together with organisms involved in nitrogen transformations and labile carbon turnover. This suggests that baseline conditions in the conventional plots before treatment application were characterised by higher moisture, micro-anoxia and fresh organic inputs.

Together, these patterns suggest an ecological trajectory from heterogeneous initial communities (Clusters 2 and 3 at t0), through a transient colonization phase (Cluster 1 at t1), towards a progressive stabilization into a conserved and resilient core (Cluster 2 from t2 onwards). The analysis of cluster-associated taxa further revealed a functional succession of microbial groups across the cropping cycle, corroborated by the temporal dynamics observed at each sampling point.

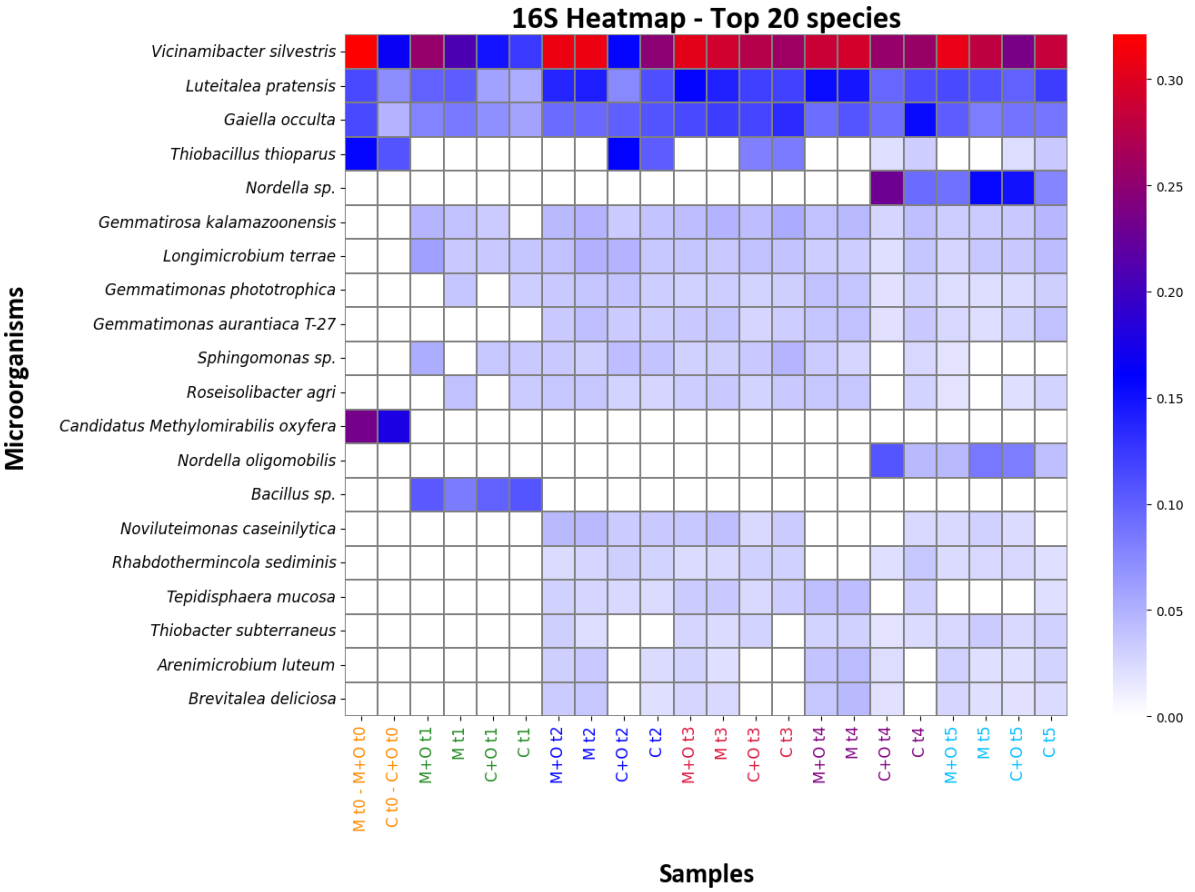


Figure 7. Heatmap of the 20 most abundant bacterial species in the 16S dataset. Rows represent bacterial species, while columns correspond to sample groups (treatments M, M+O, C and C+O) arranged along the temporal sequence of samplings (t0–t5); sample labels are colour-coded by sampling time. Cell colours indicate relative abundance, ranging from white (absence or very low abundance) through blue (intermediate values) to red (highest abundance).

The heatmap of the top 20 bacterial species (Figure 7) summarised compositional changes across treatments and time. A small set of genera remained consistently abundant throughout the experiment, forming the core microbiome, while others showed transient peaks early in the crop cycle or slight shifts associated with particular treatments. Baseline samples (t0) exhibited more pronounced differences between conventional and microbial treatment areas, whereas samples from t2 to t5 were more similar across treatments, reflecting the convergence observed in multivariate analyses. Complete profiles (relative abundance of each taxon >0.05) and details on taxa are reported in Supplementary material (Table S5).

4.Discussion

This study combined amplicon sequencing of ITS and 16S rRNA genes with unsupervised machine learning to disentangle the effects of time, crop presence and agronomic treatments on soil fungal and bacterial communities in an intensively managed greenhouse system. The results indicate that temporal dynamics and the establishment of *Valerianella locusta* were the primary drivers of community assembly, whereas the four treatments produced only subtle compositional shifts. Fungal communities

exhibited pronounced temporal restructuring and higher responsiveness to treatments, while bacterial communities converged towards a relatively stable core.

4.1 Temporal dynamics and crop presence dominate over treatment effects

Our results provide complementary insights into the temporal and functional dynamics of soil microbial communities, integrating fungal (ITS) and bacterial (16S) profiles across treatments and sampling times. A consistent pattern emerges: time, rather than treatment, is the primary driver of microbial community assembly and succession.

At t0, when *V. locusta* was not yet established, microbial assemblages reflected bulk-soil conditions, dominated by decomposer fungi (*Mortierella alpina*, *Pseudeurotium bakeri*) and metabolically specialized bacteria (*Candidatus Methylomirabilis oxyfera*, *Thiobacillus spp.*). These communities likely mirror the background soil legacy shaped by past management, organic residues, and resource availability in the absence of plant influence. Indeed, the genus *Mortierella* is well documented as a ubiquitous saprotroph enriched in bulk soils (Eberly et al. 2024), and early rhizosphere colonization is often dominated by decomposers before giving way to more specialized taxa (Fracchia et al. 2024).

From t1 onwards, the establishment of *V. locusta* introduced the "rhizosphere effect," whereby root exudates and crop residues created novel ecological niches. This shift progressively favoured antagonistic and plant-associated fungi (*Trichoderma* spp., *Epicoccum* spp.) and oligotrophic bacteria (Acidobacteriota, Actinobacteriota), consistent with previous studies (Windisch et al. 2021). The distinct *V. locusta* root bacteriome (García Méndez et al. 2023) may further contribute to these successional changes. Altogether, these dynamics point to a transition from bulk-soil decomposer communities to rhizosphere-shaped communities, progressively stabilizing along the crop growth cycle.

Despite the application of organic matter and microbial inoculants, treatment effects on community composition were modest compared with temporal changes. Several factors may explain this limited differentiation. First, the greenhouse system is characterised by highly homogeneous management (irrigation, fertilisation, environmental control) and relatively small spatial scale, which tend to reduce between-plot variation. Second, the short duration of each crop cycle and the overlap between successive cycles may not allow sufficient time for treatment-specific communities to develop and stabilise. Third, the added microorganisms and organic matter were applied according to commercial practice, which is designed primarily to enhance crop performance rather than to induce strong, lasting shifts in the entire microbiome. We note that the microbial inoculant was applied once per crop cycle, always using the same commercial formulation and recommended dose, so that our design effectively contrasted the presence versus absence of the inoculant as part of the overall management regime. Consequently, our inferences on inoculation effects are restricted to this specific product and application protocol, and the treatments are best interpreted as categorical management options rather than as quantitative levels of a continuous input signal. A dynamical-systems approach aimed at building a robust predictive model would require varying inoculant type, concentration and timing to generate richer input–output relationships and to characterise the underlying response functions more explicitly.

Under these conditions, the soil microbiota appears to be governed by strong deterministic forces linked to crop presence and overall management, while the treatments tested here act as secondary modifiers that slightly alter the prevalence of particular taxa or clusters, rather than restructuring the community as a whole. This does not exclude functional consequences of treatments, subtle shifts in key taxa may have disproportionate effects on nutrient cycling or plant health, but it suggests that such effects are embedded within a robust and recurrent temporal trajectory.

Taken together, these patterns indicate that our temporal resolution captures broad community states rather than fine-grained dynamics. For fungi, t0 represents a distinct baseline state, samples from t1 to t4 form an intermediate rhizosphere-influenced state, and t5 marks a late-stage shift in community composition. For bacteria, t0 samples form distinct baseline groups, t1 represents a short-lived transient configuration, and samples from t2 to t5 converge towards a common core community that remains relatively stable across treatments and late sampling times.

However, with six sampling times over the January–September period, we cannot resolve short-term transients, such as rapid adjustments immediately after crop establishment, inoculant application or changes in environmental conditions. In particular, the transitions between t0 and t1 for fungi and between t1 and t2 for bacteria, as well as the detailed dynamics leading to the late-season divergence observed at t5, remain under-sampled. The divergence at t5 of the fungal communities may be linked to the cumulative effects of successive crop cycles and end-of-season changes in greenhouse climate and soil conditions, but our design does not allow us to pinpoint the exact drivers. Higher-frequency sampling would be required to describe these transient phases in more detail.

4.2 Differential plasticity of fungal and bacterial communities

The functional transition observed reflects classic successional theory in soil microbial ecology: initial dominance by decomposer guilds gives way to biocontrol and plant-associated taxa linked to plant health and rhizosphere interactions (Yu et al. 2025). At t0, fungal communities were dominated by saprotrophs (*Mortierella* spp., *Pseudeurotium* spp.) linked to organic matter turnover (Eberly et al. 2024), whereas later stages (t1–t5) were enriched in antagonistic fungi (*Trichoderma* spp., *Epicoccum* spp.) with known roles in plant growth promotion and pathogen suppression (Guzmán-Guzmán et al. 2023; Yao et al. 2023).

On the bacterial side, the shift observed from opportunistic taxa and species associated with specific microenvironments (e.g., *Candidatus Methylomirabilis oxyfera*, *Pseudomonas* sp., *Thiobacillus* sp.) at t0, to oligotrophic and stress-tolerant groups (Acidobacteriota, Actinobacteriota, Gemmatimonadota) from t1 onward, is consistent with microbial life history theory (Zhang et al. 2025; Bao et al. 2021).

A central outcome of this study is the contrasting behaviour of fungi and bacteria.

The increased representation of *Trichoderma* spp. and other potential biocontrol fungi in mid- to late-stage clusters is consistent with the idea that fungal communities can respond rapidly to changes in resource inputs, root exudation patterns and management, thereby modulating plant–soil feedbacks over relatively short time scales. Several studies have shown that soil fungi are particularly sensitive and fast

responders to shifts in agricultural practices and organic matter inputs, often exhibiting stronger or earlier changes than bacterial communities following modifications in tillage, fertilisation or mulching regimes (Zheng et al. 2018; Coller et al. 2021; Wipf et al. 2021; Zhang et al. 2024). Moreover, experimental work on rhizosphere interactions has demonstrated that root exudates can rapidly attract and stimulate specific fungal groups, including *Trichoderma* spp., and that these exudate-driven shifts in fungal colonisation are associated with changes in plant defence activation and plant–soil feedback strength over a single growing season (Lombardi et al. 2018; Dutta et al. 2023; Steinauer et al. 2023). Together, these findings support our observation that fungal communities, and in particular biocontrol-related taxa such as *Trichoderma*, can track short-term changes in resource supply and rhizosphere conditions and thus act as dynamic modulators of plant–soil interactions.

In contrast, bacterial communities were dominated by a relatively stable oligotrophic core that persisted across treatments and time once the initial transient phase had passed. This pattern is consistent with the view that soil bacterial assemblages often contain a large fraction of slow-growing, oligotrophic taxa (e.g. Acidobacteriota, many Actinobacteria) that form a persistent backbone of the microbiome and support key ecosystem processes under a range of environmental conditions (Tecon and Or 2017; Huusko et al. 2024; Jaeger et al. 2024). Such persistent cores are thought to exhibit a high degree of functional redundancy and niche complementarity, whereby multiple bacterial taxa perform overlapping roles in carbon and nutrient cycling, thereby buffering soil functions against species turnover and environmental fluctuations (Yin et al. 2000; Shade et al. 2012; Miki et al. 2014; Sauma-Sánchez et al. 2024).

These differences underscore the complementary roles of fungi and bacteria in managed soils: fungi as more plastic responders that track changes in plant inputs and management, and bacteria as a resilient backbone that supports stable ecosystem processes. Recognising this complementarity is essential for designing management strategies that aim to harness the microbiome for soil health and crop productivity.

Because the present study is restricted to one greenhouse system, one crop species and a single period, these fungal–bacterial behaviour should be interpreted as site-specific and hypothesis-generating rather than universally generalisable. However, the information obtained provides a valuable baseline and reference framework for subsequent studies. Future work across multiple sites, years and crop types will be needed to assess how general these temporal trajectories and differences in plasticity between fungi and bacteria are under contrasting soil, crop and climate conditions.

4.3 Added value of unsupervised machine learning for soil microbiome analysis

The unsupervised ML pipeline combining PCA, K-means clustering and t-SNE proved effective in reducing the complexity of high-dimensional microbiome data, defining discrete microbial states and visualising temporal trajectories that are not readily captured by conventional diversity metrics alone. Dimensionality-reduction techniques, such as PCA, are widely used in microbiome research to summarise major compositional gradients, while non-linear embeddings, such as t-SNE, can better

reveal local neighbourhood structure and subtle transitions among communities. Recent reviews h ave highlighted the growing use of such unsupervised methods, including PCA, K-means and t-SNE, as core tools for exploratory analysis and feature extraction in microbiome and other omics datasets (Hernández Medina et al. 2022; Abavisani et al. 2024; Zhou and Zhao 2025). In line with these recommendations, our combined use of these techniques allowed us to collapse hundreds of taxa into a small number of ecologically interpretable axes, partition samples into recurrent microbial states associated with specific temporal phases, and visualise non-linear trajectories from baseline to late-cycle communities. These dynamics are not readily captured by conventional diversity metrics alone. Traditional α-diversity indices (e.g. richness, Shannon) compress each community into a single value, so communities with markedly different taxonomic and functional profiles can display similar diversity scores (Finotello et al. 2018; Cassol et al. 2025). Likewise, standard β-diversity measures and their ordinations (e.g. Bray–Curtis distances visualised by PCoA) summarise overall compositional dissimilarity but do not explicitly identify recurrent community states or non-linear transitions over time, and their sensitivity depends on the chosen metric and normalisation strategy (Lemos et al. 2011; Kers and Saccenti 2022; Bars-Cortina 2022). Importantly, the ML-based clusters were ecologically interpretable: in fungi, they aligned with successional stages from decomposer-dominated to rhizosphere-influenced communities; in bacteria, they captured the progression from distinct baseline assemblies to a convergent core.

These observations are in line with recent reviews and applications showing that dimensionality reduction and clustering algorithms are increasingly used to extract latent structure from microbiome and other omics datasets, going beyond classic α/β-diversity summaries and providing a pattern-oriented foundation for downstream predictive modelling (Zhou and Gallins 2019; Shi et al. 2022; Armstrong et al. 2022; Abavisani et al. 2024).

5.Conclusions

By combining metagenomic amplicon sequencing of ITS and 16S rRNA genes with multivariate and unsupervised machine learning analyses, this study clarified how soil fungal and bacterial communities respond to time, crop presence and agronomic treatments in an intensively managed greenhouse system for *Valerianella locusta*. Across both datasets, temporal dynamics and rhizosphere effects emerged as the dominant drivers of community assembly, whereas the four treatments (C, C+O, M, M+O) produced only modest compositional shifts.

Fungal communities exhibited marked plasticity, transitioning from saprotroph-dominated assemblages at baseline to communities enriched in antagonistic and plant-associated taxa at later stages. Bacterial communities, in contrast, rapidly converged towards a stable oligotrophic core dominated by Acidobacteriota, Actinobacteriota and other functionally redundant taxa, with only a brief transient phase following crop establishment. These contrasting responses highlight the complementary roles of fungi and bacteria in managed soils and the different degrees of stability and responsiveness they confer to the system.

The unsupervised ML analyses synthesised these patterns into a small number of recurrent microbial "states" and temporal trajectories, offering an integrated view of microbiome dynamics that complements conventional diversity metrics without replacing them.

From an applied perspective, our findings indicate that, in intensive greenhouse baby-leaf production, strategies aimed at improving soil bio-sustainability should explicitly account for the strong and recurrent imprint of time and crop presence on the microbiome. Numerous studies have shown that plant development and rhizosphere processes can exert a stronger and more consistent influence on microbial community structure than individual management interventions, especially in systems with homogeneous inputs and short crop cycles (Schmidt et al. 2019; Wang et al. 2022; Zhao et al. 2019; Han et al. 2025). In this context, the modest treatment effects observed here suggest that management practices such as organic amendments and microbial inoculants may operate mainly by modulating specific guilds within a robust temporal trajectory, rather than by reshaping the community as a whole.

Our results also support the idea that fungal and bacterial communities contribute differently to soil functioning and respond with different degrees of plasticity and stability. Recent work has shown that fungal networks often display higher structural stability or stronger biogeographical structuring than bacterial networks, whereas bacteria can be more resilient and functionally redundant under environmental change (de Vries et al. 2018; Gschwend et al. 2021; Li et al. 2024; Wu et al. 2025; Fu et al. 2024). In our system, fungi behaved as more responsive levers of change, tracking shifts in resource inputs and crop development, while bacteria converged towards a stable oligotrophic core. This suggests that management strategies aiming to enhance soil bio-sustainability in greenhouse baby-leaf systems may be more effective if they explicitly target fungal functions, for example by promoting beneficial saprotrophic and antagonistic taxa, while preserving the stability of the bacterial backbone that underpins key biogeochemical processes.

While the present work is exploratory and does not implement predictive models, the identification of recurrent microbial states and their associated taxa represent a crucial step towards AI-assisted soil health assessment (Woodman and Mangoni 2023; Xu et al. 2020). In future work, these states and their discriminant taxa could be used as features in supervised ML models linking microbiome composition to soil functions, crop performance or resilience to stress, thereby advancing the development of predictive tools for bio-sustainable management.

Author contributions: Conceptualisation and methodology: R.P., M.M.M. and M.R.; formal analysis and data curation: R.P.; visualization, validation and supervision: R.P., M.M.M. and M.R.; writing—original draft preparation: R.P.; writing—review and editing: R.P., M.M.M., A.A., S.C. and M.R.; funding acquisition: M.M.M and M.R. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This study was carried out within a PhD funded by EVJA s.r.l.

Funding: This study was carried out within the Agritech National Research Center and received funding from the European Union Next-Generation EU (Piano Nazionale di Ripresa e Resilienza (PNRR) – missione 4 componente 2, investimento 1.4 – d.d. 1032 17/06/2022, cn00000022). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Conflict of Interest: The authors declare that they have no conflict of interest.

Data availability statement: The metagenomic sequencing data that support the findings of this study have not yet been deposited in a public repository but are available from the corresponding author upon reasonable request, and will be deposited in an appropriate public sequence repository no later than April 2026.

Reference list

Abavisani M, Khoshrou A, Foroushan SK, Ebadpour N, Sahebkar A (2024) Deciphering the gut microbiome: The revolution of artificial intelligence in microbiota analysis and intervention. Current Research in Biotechnology, 7, 100211.

Armstrong G, Rahman G, Martino C, McDonald D, Gonzalez A, Mishne G, Knight R (2022) Applications and comparison of dimensionality reduction methods for microbiome data. Frontiers in bioinformatics, 2, 821861.

Bao Y, Dolfing J, Guo Z, Chen R, Wu M, Li Z, Lin X, Feng Y (2021) Important ecophysiological roles of non-dominant Actinobacteria in plant residue decomposition, especially in less fertile soils. Microbiome 9: 84.

Bars-Cortina D (2022) Alpha and beta-diversities performance comparison between different normalization methods and centered log-ratio transformation in a microbiome public dataset. bioRxiv, 2022-11.

Bhaduri D, Sihi D, Bhowmik A, Verma BC, Munda S, Dari B (2022) A review on effective soil health bio-indicators for ecosystem restoration and sustainability. Front Microbiol 13: 938481.

Burrill HM, Magnoli SM, Bever JD (2025) Soil microbiome composition is highly responsive to precipitation and plant composition manipulations in a field biodiversity experiment. Front Microbiomes 4.

Cassol I, Ibañez M, Bustamante JP (2025) Key features and guidelines for the application of microbial alpha diversity metrics. Scientific Reports, 15(1), 622.

Cole L, Goodall T, Jehmlich N, Griffiths RI, Gleixner G, Gubry-Rangin C, Malik AA (2024) Land use effects on soil microbiome composition and traits with consequences for soil carbon cycling. ISME Commun 4: 000000.

Coller E, Oliveira Longa CM, Morelli R, Zanoni S, Cersosimo Ippolito MC, Pindo M, Cappelletti C, Ciutti F, Menta C, Zanzotti R, Ioriatti C (2021) Soil communities: who responds and how quickly to a change in agricultural system? Sustainability 14: 383.

De Vries FT, Griffiths RI, Bailey M, Craig H, Girlanda M, Gweon HS, Hallin S, Kaisermann A, Keith AM, Kretzschmar M, Lemanceau P, Lumini E, Mason KE, Oliver A, Ostle N, Prosser JI, Thion C, Thomson B, Bardgett RD (2018) Soil bacterial networks are less stable under drought than fungal networks. Nat Commun 9: 3033.

Doran JW, Zeiss MR (2000) Soil health and sustainability: managing the biotic component of soil quality. Appl Soil Ecol 15: 3–11.

Dutta P, Mahanta M, Singh SB, Thakuria D, Deb L, Kumari A, Upamanya GK, Boruah S, Dey U, Mishra AK, Vanlaltani L, VijayReddy D, Heisnam P, Pandey AK (2023) Molecular interaction between plants and Trichoderma species against soil-borne plant pathogens. Front Plant Sci 14: 1145715.

Eberly JO, Hurd A, Oli D, Dyer AT, Seipel TF, Carr PM (2024) Compositional profiling of the rhizosphere microbiome of Canada thistle reveals consistent patterns across the United States northern Great Plains. Sci Rep 14: 18016.

Ferris H, Tuomisto H (2015) Unearthing the role of biological diversity in soil health. Soil Biol Biochem 85: 101–109.

Fierer N, Wood SA, Bueno de Mesquita CP (2021) How microbes can, and cannot, be used to assess soil health. Soil Biol Biochem 153: 108111.

Finotello F, Mastrorilli E, Di Camillo B (2018) Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in bioinformatics, 19(4), 679-692.

Fracchia F, Guinet F, Engle NL, Tschaplinski TJ, Veneault-Fourrey C, Deveau A (2024) Microbial colonisation rewires the composition and content of poplar root exudates, root and shoot metabolomes. Microbiome 12: 173.

Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ (2008) Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. Appl Environ Microbiol 74: 2461–2470.

Fu X, Dai Y, Cui J, Deng P, Fan W, Xu X (2024) Soil bacterial and fungal communities resilience to long-term nitrogen addition in subtropical forests in China. J For Res 35: 17.

García Méndez S, Persyn A, De Tender C, Goormachtig S, Willems A (2023) Unravelling the bacterial community composition of Valerianella locusta, a cold-tolerant plant. Phytobiomes J 7: 500–514.

Gschwend F, Hartmann M, Hug AS, Enkerli J, Gubler A, Frey B, Meuli RG, Widmer F (2021) Long-term stability of soil bacterial and fungal community structures revealed in their abundant and rare fractions. Mol Ecol 30: 4305–4320.

Guzmán-Guzmán P, Kumar A, de los Santos-Villalobos S, Parra-Cota FI, Orozco-Mosqueda MdC, Fadiji AE, Hyder S, Babalola OO, Santoyo G (2023) Trichoderma species: our best fungal allies in the biocontrol of plant diseases—a review. Plants 12: 432.

Han M, Han Y, Liu X, Li G, Li P (2025) Rhizosphere microbiomes altered by environmental stresses and agronomic practices: Implications for plant adaptation and soil biogeochemical processes. Plant Stress, 101062.

Hartmann M, Six J (2023) Soil structure and microbiome functions in agroecosystems. Nat Rev Earth Environ 4: 4–18.

He Z, Yuan C, Chen P, Rong Z, Peng T, Farooq TH, Wang G, Yan W, Wang J (2023) Soil microbial community composition and diversity analysis under different land use patterns in Taojia River Basin. Forests 14: 000000.

Hernández Medina R, Kutuzova S, Nielsen KN, Johansen J, Hansen LH, Nielsen M, Rasmussen S (2022) Machine learning and deep learning applications in microbiome research. ISME Commun 2:98.

Hossain MA, Egidi E, Xiong C, Singh BK (2025) Soil multitrophic interactions in a changing world. J Sustain Agric Environ 4: 000000.

Hou S, Wolinska KW, Hacquard S (2021) Microbiota-root-shoot-environment axis and stress tolerance in plants. Curr Opin Plant Biol 62: 102028.

Huusko K, Manninen OH, Myrsky E, Stark S (2024) Soil fungal and bacterial communities reflect differently tundra vegetation state transitions and soil physico-chemical properties. New Phytol 243: 407–422.

Jaeger ACH, Hartmann M, Conz RF, Six J, Solly EF (2024) Prolonged water limitation shifts the soil microbiome from copiotrophic to oligotrophic lifestyles in Scots pine mesocosms. Environ Microbiol Rep 16: e13211.

Jung S, Dagobert T, Morel JM, Facciolo G (2024) A review of t-SNE. Image Process Online 14: 250–270.

Kapoor D, Sharma P, Sharma MMM, Yadav S, Husen A (2024) Exploring soil microbiota and their role in plant growth, stress tolerance, disease control and nutrient immobilizer. Biocatal Agric Biotechnol 61: 103358.

Kers JG, Saccenti E (2022) The power of microbiome studies: some considerations on which alpha and beta metrics to use and how to report results. Frontiers in microbiology, 12, 796025.

Kopittke PM, Minasny B, Pendall E, Rumpel C, McKenna BA (2024) Healthy soil for healthy humans and a healthy planet. Crit Rev Environ Sci Technol 54: 210–221.

Kuzyakov Y, Razavi BS (2019) Rhizosphere size and shape: Temporal dynamics and spatial stationarity. Soil Biol Biochem 135: 343–360.

Lehman RM, Acosta-Martinez V, Buyer JS, Cambardella CA, Collins HP, Ducey TF, Halvorson JJ, Jin VL, Johnson JMF, Kremer RJ, Lundgren JG, Manter DK, Maul JE, Smith JL, Stott DE (2015) Soil biology for resilient, healthy soil. J Soil Water Conserv 70: 12A–18A.

Lehmann J, Kleber M (2015) The contentious nature of soil organic matter. Nature 528: 60–68.

Lemos LN, Fulthorpe RR, Triplett EW, Roesch LF (2011) Rethinking microbial diversity analysis in the high throughput sequencing era. Journal of microbiological methods, 86(1), 42-51.

Li D, Wu C, Wu J (2024) Soil fungal community has higher network stability than bacterial community in response to warming and nitrogen addition in a subtropical primary forest. Appl Environ Microbiol 90: e00001-24.

Lombardi N, Vitale S, Turrà D, Reverberi M, Fanelli C, Vinale F, Marra R, Ruocco M, Pascale A, D'Errico G, Woo SL, Lorito M (2018) Root exudates of stressed plants stimulate and attract Trichoderma soil fungi. Mol Plant Microbe Interact 31: 982–994.

Martinez CM, Alvarez LH, Celis LB, Cervantes FJ (2013) Humus-reducing microorganisms and their valuable contribution in environmental processes. Appl Microbiol Biotechnol 97: 10293–10308.

Masuda Y, Mise K, Xu Z, Zhang Z, Shiratori Y, Senoo K, Itoh H (2024) Global soil metagenomics reveals distribution and predominance of Deltaproteobacteria in nitrogen-fixing microbiome. Microbiome 12: 000000.

Miki T, Yokokawa T, Matsui K (2014) Biodiversity and multifunctionality in a microbial community: a novel theoretical approach to quantify functional redundancy. Proc R Soc B 281: 20132498.

Mirhendi H, Diba K, Rezaei A, Jalalizand N, Hosseinpur L, Khodadadi H (2007) Colony-PCR is a rapid and sensitive method for DNA amplification in yeasts. Iran J Public Health 36: 40–45.

Mueller CW, Baumert V, Carminati A, Germon A, Holz M, Kögel-Knabner I, Peth S, Schlüter S, Uteau D, Vetterlein D, Teixeira P, Vidal A (2024) From rhizosphere to detritusphere – soil structure formation driven by plant roots and the interactions with soil biota. Soil Biol Biochem 193: 108093.

Nannipieri P (2020) Soil is still an unknown biological system. Appl Sci 10: 2324.

Omae N, Tsuda K (2022) Plant–microbiota interactions in abiotic stress environments. Mol Plant Microbe Interact 35: 511–526.

Omotayo OP, Babalola OO (2021) Resident rhizosphere microbiome's ecological dynamics and conservation: Towards achieving the envisioned sustainable development goals, a review. Int Soil Water Conserv Res 9: 127–142.

Pace R, Schiano Di Cola V, Monti MM, Affinito A, Cuomo S, Loreto F, Ruocco M (2025) Artificial intelligence in soil microbiome analysis: a potential application in predicting and enhancing soil health—a review. Discov Appl Sci 7: 000000.

Pandey A, Singh SK, Yadav B, Mukherjee S (2024) Rhizospheric microorganisms for the sustainable management of plant health. Sustain Agric Rev 52: 45–72.

Pedrinho A, Mendes LW, de Araujo Pereira AP, Araujo ASF, Vaishnav A, Karpouzas DG, Singh BK (2024) Soil microbial diversity plays an important role in resisting and restoring degraded ecosystems. Plant Soil 500: 325–349.

Sauma-Sánchez T, Alcorta J, Tamayo-Leiva J, Díez B, Bezuidenhout H, Cowan DA, Ramond JB (2024) Functional redundancy buffers the effect of poly-extreme environmental conditions on southern African dryland soil microbial communities. FEMS Microbiol Ecol 100: fiae157.

Schmidt JE, Kent AD, Brisson VL, Gaudin AC (2019) Agricultural management and plant selection interactively affect rhizosphere microbial community structure and nitrogen cycling. Microbiome 7: 146.

Shade A, Peter H, Allison SD, Baho DL, Berga M, Bürgmann H, Huber DH, Langenheder S, Lennon JT, Martiny JBH, Matulich KL, Schmidt TM, Handelsman J (2012) Fundamentals of microbial community resistance and resilience. Front Microbiol 3: 417.

Shi Y, Zhang L, Peterson CB, Do KA, Jenq RR (2022) Performance determinants of unsupervised clustering methods for microbiome data. Microbiome 10: 25.

Steinauer K, Thakur MP, Hannula SE, Weinhold A, Uthe H, van Dam NM, Bezemer TM (2023) Root exudates and rhizosphere microbiomes jointly determine temporal shifts in plant–soil feedbacks. Plant Cell Environ 46: 1885–1899.

Taraboletti A, King A, Dixon Y, Orr O, Parnell C, Watson Y, Nash B, Esimai C, Ude G (2023) Assessing microbial diversity in soil samples along the Potomac River: implications for environmental health. Microbiol Spectr 11: e04566-22.

Tecon R, Or D (2017) Biophysical processes supporting the diversity of microbial life in soil. FEMS Microbiol Rev 41: 599–623.

Wang P, Yan S, Zhang W, Xie X, Li M, Ren T, Cong R, Zhang Z (2022) Effect of soil management systems on the rhizosphere bacterial community structure of tobacco: Continuous cropping vs. paddy-upland rotation. Front Plant Sci 13: 996858.

Wang YF, Xu JY, Liu ZL, Cui HL, Chen P, Cai TG, Li G, Ding LJ, Qiao M, Zhu YG, Zhu D (2024) Biological interactions mediate soil functions by altering rare microbial communities. Environ Sci Technol 58: 5866–5877.

Windisch S, Sommermann L, Babin D, Chowdhury SP, Grosch R, Moradtalab N, Walker F, Höglinger B, El-Hasan A, Armbruster W, Nesme J, Sørensen SJ, Schellenberg I, Geistlinger J, Smalla K, Rothballer M, Ludewig U, Neumann G (2021) Impact of long-term organic and mineral fertilization on rhizosphere metabolites, root–microbial interactions and plant health of lettuce. Front Microbiol 11: 597745.

Wipf HML, Xu L, Gao C, Spinner HB, Taylor J, Lemaux P, Mitchell J, Coleman-Derr D (2021) Agricultural soil management practices differentially shape the bacterial and fungal microbiomes of Sorghum bicolor. Appl Environ Microbiol 87: e02345-20.

Woodman RJ, Mangoni AA (2023) A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future. Aging Clin Exp Res 35: 2363–2397.

Wu Y, Jia Y, Liu Q, Guo J, Shen Y, Yu L, Zhang G, Deng J, Tian H, Qi L (2025) Bacterial and fungal communities exhibit contrasting assembly processes during native bamboo expansion through coupled vegetation–soil modifications across elevation gradients. Glob Ecol Conserv 62: e03822.

Xu X, Xie Z, Yang Z, Li D, Xu X (2020) A t-SNE based classification approach to compositional microbiome data. Front Genet 11: 620143.

Yao X, Guo H, Zhang K, Zhao M, Ruan J, Chen J (2023) Trichoderma and its role in biological control of plant fungal and nematode disease. Front Microbiol 14: 1160551.

Yin B, Crowley D, Sparovek G, De Melo WJ, Borneman J (2000) Bacterial functional redundancy along a soil reclamation gradient. Appl Environ Microbiol 66: 4361–4365.

Yu Z, Zeng X, Cheng X, Zhang Q, Zhang K (2025) Patterns and environmental drivers of soil microbial succession. Glob Chang Biol 31: 000000.

Yuan C, Yang H (2019) Research on K-value selection method of K-means clustering algorithm. J (Basel) 2: 226–235.

Zhang Y, Huang J, Xue J, Zhang K, Chen X, Jia J, Huang Q (2025) Characterization of soil bacterial communities in different vegetation types on the lava plateau of Jingpo Lake. Microorganisms 13: 1648.

Zhang L, Pei N, Wu F, Wu N, Li Z, Huang Y, Tan M, Yang Y (2024) Short-term response of soil fungal community composition and trophic mode to winter mulch management in Lei bamboo (Phyllostachys praecox) forests. Appl Soil Ecol 201: 105526.

Zhao M, Yuan J, Shen Z, Dong M, Liu H, Wen T, Li R, Shen Q (2019) Predominance of soil vs root effect in rhizosphere microbiota reassembly. FEMS Microbiol Ecol 95: fiz139.

Zheng W, Zhao Z, Gong Q, Zhai B, Li Z (2018) Responses of fungal–bacterial community and network to organic inputs vary among different spatial habitats in soil. Soil Biol Biochem 125: 54–63.

Zhou YH, Gallins P (2019) A review and tutorial of machine learning methods for microbiome host trait prediction. Front Genet 10: 579.

Zhou T, Zhao F (2025) AI-empowered human microbiome research. Gut, in press.

Zong D, Zhou Y, Zhou J, Zhao Y, Hu X, Wang T (2024) Soil microbial community composition by crop type under rotation diversification. BMC Microbiol 24: 000000.

3.3 Supplementary materials

SUPPLEMENTARY MATERIAL - S1

Preliminary data cleaning and exploratory analysis were performed using Python (v3.11) as described below:

- *pandas* (v2.0.3) and *numpy* (v1.25.0) for data handling and transformation;
- *scikit-learn* (v1.3.0) for dimensionality reduction and clustering;
- *scipy* (v1.11.1) for statistical calculations;
- *matplotlib* (v3.7.2) and *seaborn* (v0.12.2) for data visualization.

The analysis pipeline included:

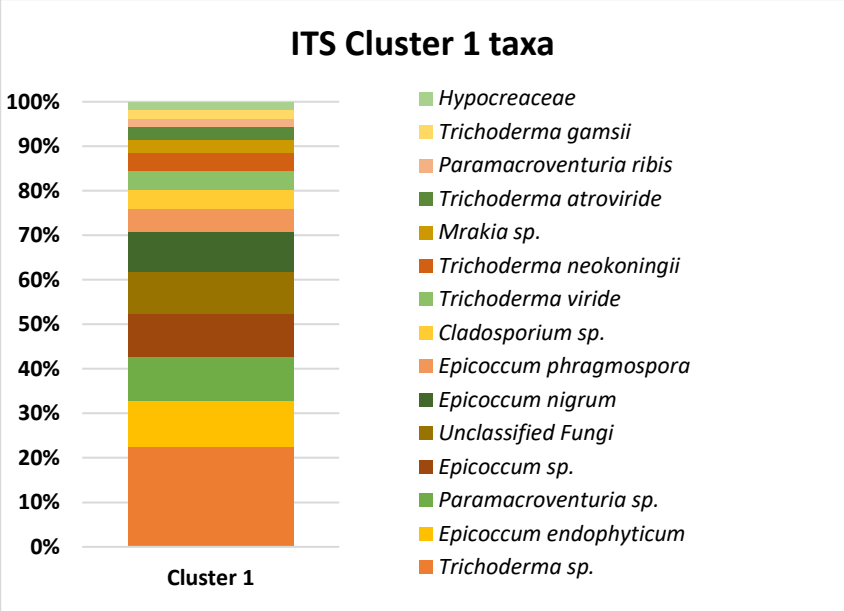
- Data Preprocessing:
  Missing values (NaN) were replaced with zeros. Taxa with zero abundance across all samples were excluded;
- Taxonomic Aggregation and Cleaning:
  Identical taxa were merged by summing their values across samples;
- Replicate Aggregation and Normalization:
  Biological replicates (n=3) were averaged. Total Sum Scaling (TSS) was applied to obtain relative abundances;
- Filtering by Abundance:
  Taxon with relative abundance ≥ 0.5% in at least one sample were retained;
- Ordination and Clustering:
  Heatmaps were generated with hierarchical clustering based on Bray-Curtis dissimilarity. PCA was performed for ordination and dimensionality reduction;
- Advanced Pattern Detection:
  K-means clustering was applied to identify sample groupings. The optimal number of clusters was first estimated using the silhouette score, and subsequently refined by considering their biological interpretability;
- Cluster-level taxonomic signatures:
  To characterise the taxa associated with each of the three K-means clusters, we combined the cluster assignments with the taxon relative-abundance tables. For each dataset (ITS and 16S), we calculated the mean relative abundance of each taxon within each cluster and used these values to identify the taxa that were most abundant and characteristic of a given cluster.
- Finally, t-SNE was employed for nonlinear dimensionality reduction and visual exploration of treatment-specific patterns.

SUPPLEMENTARY MATERIAL – S2

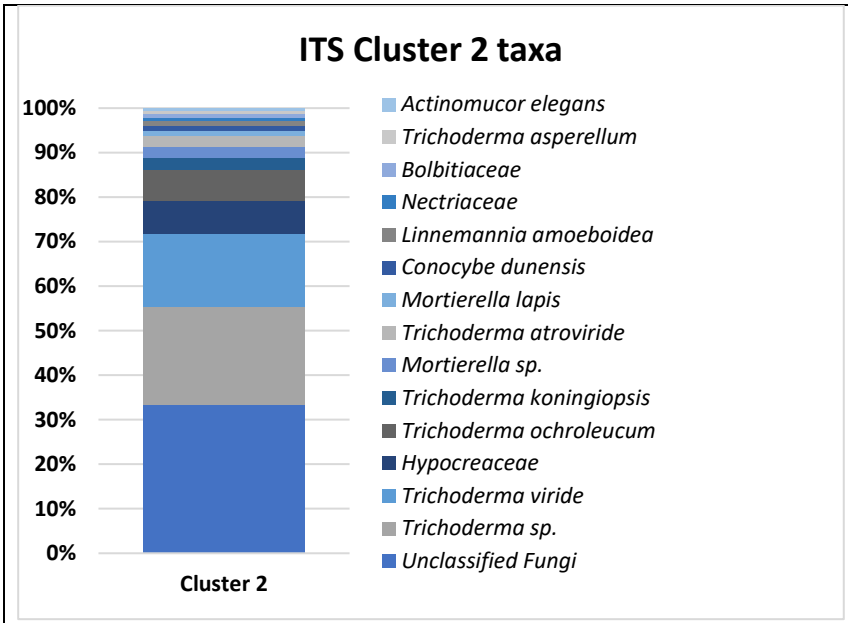
Cluster-level taxonomic signatures

Microorganisms	Cluster 1	Cluster 2	Cluster 3
Actinomucor elegans	0	0,00590626	0,016949153
Actinomucor elegans var. kuwaitensis	0,001179245	0	0
Ascomycota	0,009283036	0,003734107	0
Aureobasidium pullulans	0	0,000781746	0
Basidiomycota	0	0,005329599	0
Bolbitiaceae	0	0,006909918	0
Botryotrichum atrogriseum	0	0,000470905	0
Cladosporiaceae	0,010786071	0	0
Cladosporium halotolerans	0	0,001462079	0
Cladosporium sp.	0,038778297	0,001402539	0
Conocybe ceracea	0	0,001092732	0
Conocybe dunensis	0	0,011570276	0
Conocybe fuscimarginata	0	0,001211699	0
Conocybe incarnata	0	0,002813267	0
Conocybe sp.	0	0,004561471	0
Coprinellus radians	0	0,002764865	0
Coprinellus verrucispermus	0	0	0,031548541
Didymella sp.	0,004551071	0	0
Epicoccum brasiliense	0,004514645	0	0
Epicoccum endophyticum	0,090525931	0	0
Epicoccum nigrum	0,082116321	0	0,013317191
Epicoccum phragmospora	0,044785498	0	0
Epicoccum sp.	0,087522132	0	0
Hypocreaceae	0,01621748	0,070145852	0
Linnemannia amoeboidea	0,00166762	0,008864913	0
Linnemannia elongata	0	0	0,043377847
Linnemannia exigua	0	0,000555179	0,016620499
Linnemannia gamsii	0	0,002756488	0,015697138
Mortierella alpina	0	0	0,366792315
Mortierella globalpina	0,001429388	0,002786048	0
Mortierella lapis	0,001905851	0,012597342	0
Mortierella sp.	0	0,023903397	0

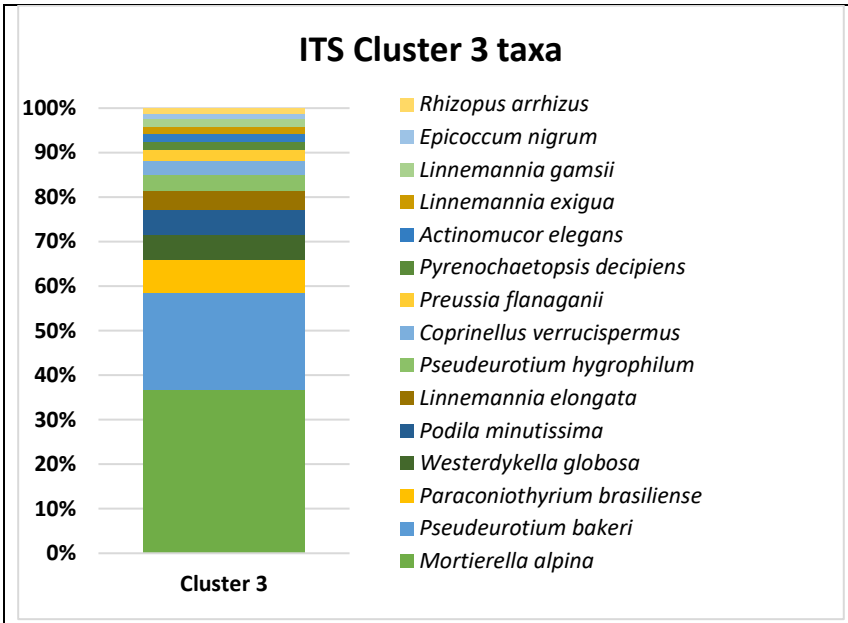
Mortierellaceae	0	0,001390809	0
Mrakia sp.	0,025600581	0	0
Nectriaceae	0	0,007751628	0
Neodidymelliopsis sp.	0,004532852	0	0
Nothophoma acaciae	0,007434235	0	0
Nothophoma sp.	0,00414891	0	0
Paraconiothyrium brasiliense	0	0	0,07180753
Paramacroventuria ribis	0,018236516	0	0
Paramacroventuria sp.	0,088597967	0	0
Pezizales	0	0,002174662	0
Plectosphaerella cucumerina	0	0,002776576	0
Podila minutissima	0	0	0,055936004
Preussia flanaganii	0	0	0,025608177
Pseudeurotium bakeri	0,005975457	0	0,219680557
Pseudeurotium hygrophilum	0	0,004093781	0,034780305
Pseudeurotium sp.	0	0,000831737	0
Pyrenochaetopsis decipiens	0	0	0,018159806
Rhizopus arrhizus	0	0,000822043	0,012003693
Saccharomyces cerevisiae	0,001155422	0,000976304	0
Sporormiaceae	0	0,002549321	0
Trichoderma asperellum	0	0,00658254	0
Trichoderma atroviride	0,024844071	0,021524598	0
Trichoderma austrokoningii	0	0,002537781	0
Trichoderma caribbaeum	0,008751894	0	0
Trichoderma dorotheae	0,0030677	0	0
Trichoderma gamsii	0,017052491	0	0
Trichoderma koningiopsis	0,005861405	0,024630221	0
Trichoderma neokoningii	0,036887054	0	0
Trichoderma ochroleucum	0,015960389	0,066938092	0
Trichoderma paucisporum	0	0,000685119	0
Trichoderma scalesiae	0,004283007	0	0
Trichoderma sp.	0,202682555	0,209919325	0
Trichoderma viride	0,037908075	0,155970037	0
Unclassified Fungi	0,083395715	0,316224746	0
Vacuiphoma ferulae	0,008361119	0	0
Westerdykella globosa	0	0	0,057721243



- Cluster 1 (C, C+O at t2 and M at t5) was characterized by a strong prevalence of *Trichoderma* spp. (including *T. viride, T. neokoningii, T. atroviride,* and *T. gamsii*) together with several species of *Epicoccum* and *Paramacroventuria*. This cluster is characterized by a dual structure: beneficial fungi (*Trichoderma* spp., *Epicoccum* spp.) co-occur with the phytopathogen *Paramacroventuria* sp. at high levels. Minor taxa such as *Cladosporium* spp., *Linnemannia amoeboidea* and *Mortierella* spp. are present but not dominant. This suggests a transitional or stressed community, where beneficial fungi and pathogens compete strongly. This may represent dynamic equilibrium under biotic stress, with potential implications for disease suppression vs. pathogen establishment.



- Cluster 2 (which includes most samples from time points t1–t4) was overwhelmingly dominated by *Trichoderma* spp. (>45%), including *T. viride, T. ochroleucum, T. koningiopsis, T. atroviride,* and *T. asperellum*. It was further characterized by a high proportion of unclassified fungi and the co-occurrence of fast-growing saprophytes such as *Mortierella* spp., *Linnemannia amoeboidea,* and *Actinomucor elegans*. This assemblage suggests a transitional stage in which antagonistic taxa coexist with opportunistic decomposers, likely reflecting shifts in resource availability or disturbance. The additional presence of saprophytic basidiomycetes (e.g., Bolbitiaceae, *Conocybe* spp., and *Coprinellus* spp.) points to active decomposition processes. Functionally, this profile aligns with rhizosphere or management conditions that promote biocontrol-oriented guilds, such as fertilization or microbial inoculation, while still sustaining a background of saprotrophic fungi.



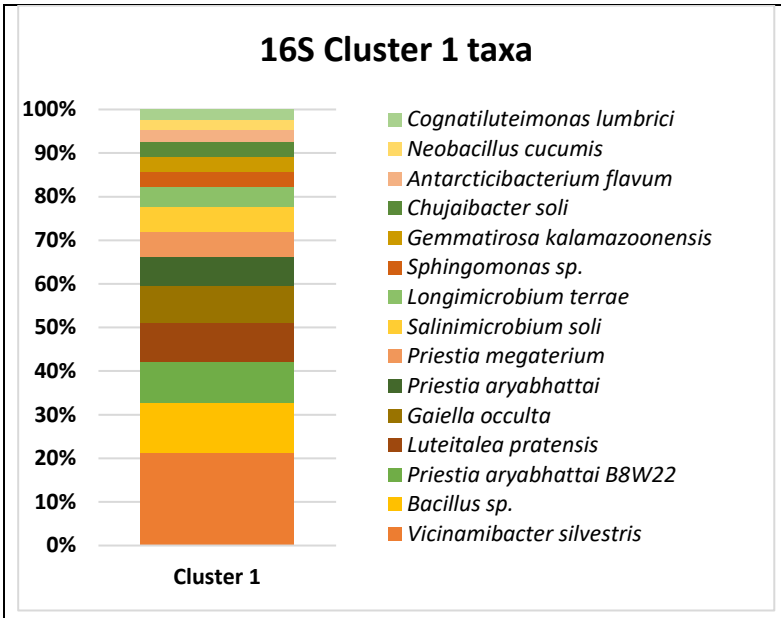
- Cluster 3 was clearly distinct (all samples at t0), dominated by *Mortierella alpina* and *Pseudeurotium bakeri*, alongside other saprophytic fungi (*Coprinellus verrucispermus*, *Linnemannia* spp., *Paraconiothyrium brasiliense*, *Podila minutissima*, *Westerdykella globosa*), pointing to a community strongly oriented towards organic matter decomposition and nutrient turnover. This may reflect a decomposer-driven cluster, suggesting an environment rich in organic residues and undergoing active carbon turnover and nutrient mineralization.

SUPPLEMENTARY MATERIAL – S3

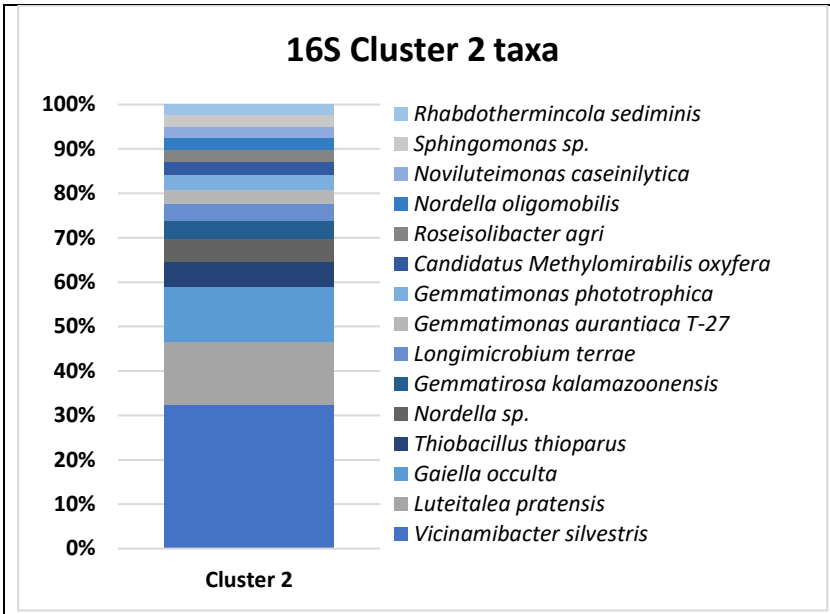
Cluster-level taxonomic signatures

microorganisms	Cluster 1	Cluster 2	Cluster 3
Aciditerrimonas ferrireducens JCM 15389	0	0,002133474	0
Antarcticibacterium flavum	0,024353599	0	0
Arenimicrobium luteum	0	0,018917546	0
Bacillus badius	0	0	0,036363636
Bacillus sp.	0,098676659	0	0
Brevitalea deliciosa	0	0,018233334	0
Brevitalea sp.	0	0,002701503	0
Candidatus Methylomirabilis oxyfera	0	0,02603526	0,177777778
Candidatus Udaeobacter copiosus	0	0,006560066	0
Chujaibacter soli	0,027794085	0	0
Cognatiluteimonas lumbrici	0,020006277	0	0
Cytobacillus spp.	0,017771601	0	0
Gaiella occulta	0,07262389	0,106721181	0,048484848
Gemmatimonas aurantiaca T-27	0	0,028595058	0
Gemmatimonas phototrophica	0,017007006	0,026895734	0
Gemmatirosa kalamazoonensis	0,030084012	0,035964557	0
Geoalkalibacter ferrihydriticus	0	0,000907153	0
Limisphaera ngatamarikiensis	0	0,003057825	0
Longimicrobium terrae	0,041452627	0,032345426	0
Luteitalea pratensis	0,078557682	0,121499078	0,072727273
Mesobacillus subterraneus	0,007024996	0	0
Neobacillus cucumis	0,02003269	0	0
Neobacillus niacini	0	0	0,034343434
Nitrosospira multiformis	0	0,001609982	0
Nocardioides sp.	0,016355057	0	0
Nordella oligomobilis	0	0,022529959	0
Nordella sp.	0	0,04412057	0
Noviluteimonas caseinilytica	0	0,02185564	0
Novilysobacter spongiicola	0,016142384	0	0
Ohtaekwangia koreensis	0	0,006594398	0
Peribacillus asahii	0	0	0,094949495

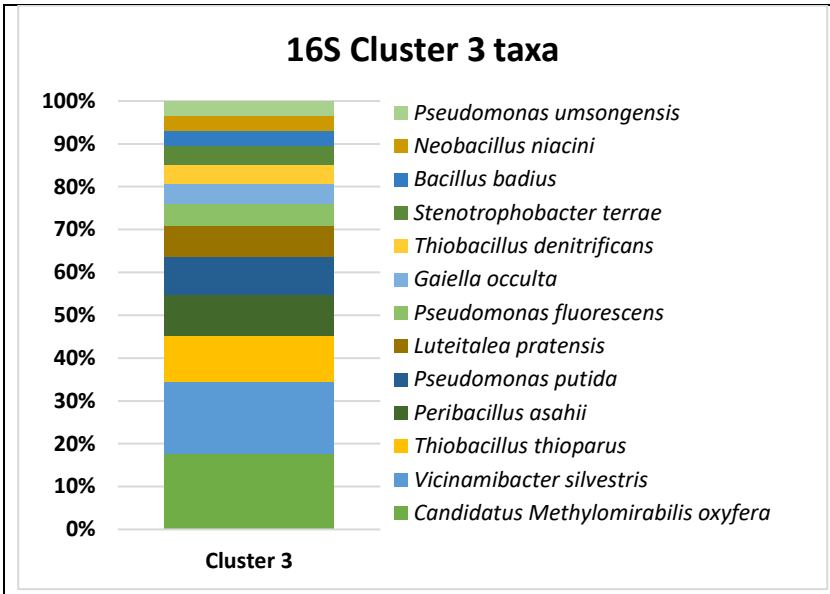
Povalibacter uvarum	0	0,012555066	0
Priestia aryabhattai	0,05614783	0	0
Priestia aryabhattai B8W22	0,081034186	0	0
Priestia megaterium	0,051329887	0	0
Priestia sp.	0,018153031	0	0
Pseudomonas fluorescens	0	0	0,050505051
Pseudomonas putida	0	0	0,088888889
Pseudomonas sp.	0	0,001252208	0
Pseudomonas umsongensis	0	0	0,034343434
Rhabdothermincola sediminis	0	0,020382897	0
Roseisolibacter agri	0,017954807	0,023448848	0
Rossellomorea marisflavi	0,008827376	0	0
Salinimicrobium soli	0,047358045	0	0
Sphingomonas sediminicola	0	0,001212313	0
Sphingomonas sp.	0,030540284	0,021754013	0
Stenotrophobacter terrae	0	0	0,042424242
Tepidisphaera mucosa	0	0,020105097	0
Thermoanaerobaculum aquaticum	0	0,002178897	0
Thiobacillus denitrificans	0	0,001498178	0,044444444
Thiobacillus sp.	0	0,007582528	0
Thiobacillus thioparus	0	0,04718986	0,109090909
Thiobacillus thiophilus	0	0,010568425	0
Thiobacter subterraneus	0	0,019072218	0
Tumebacillus ginsengisoli	0,008291129	0	0
Vicinamibacter silvestris	0,183788805	0,277194615	0,165656566
Vicinamibacter sp.	0	0,006727093	0
Vreelandella sp.	0,008692053	0	0



- Cluster 1 (t1 samples) in addition to the core taxa, shows a strong enrichment of Firmicutes/Bacillaceae (*Bacillus sp.*, *Priestia aryabhattai*, *P. megaterium*, *Neobacillus cucumis*) and versatile/tolerant taxa (*Salinimicrobium soli*, *Longimicrobium terrae*, *Sphingomonas sp.*, *Gemmatirosa kalamazoonensis*). This profile is consistent with more aerated conditions and moderate organic inputs: spore-formers suggest disturbance/drying cycles and persistence capacity; *Sphingomonas* points to the degradation of complex/aromatic compounds. The presence of *Salinimicrobium* also indicates neutral to alkaline soils with moderate osmotic stress.



- Cluster 2 (M and M+O t0 e all samples from t2 to t5), alongside the shared core, includes *Thiobacillus thioparus* (thiosulfate/sulfur oxidation), members of Gemmatimonadota (*Gemmatirosa kalamazoonensis*, *Gemmatimonas phototrophica*, *G. aurantiaca*), *Nordella* spp., *Roseisolibacter agri*, *Noviluteimonas caseinilytica*, *Rhabdothermincola sediminis*, and *Sphingomonas*. Altogether, these indicate micro-niches with reduced sulfur compounds and the presence of Gemmatimonadota, often linked to drier/oxidized soils. Cluster 2 therefore appears intermediate/transitional: it maintains an oligotrophic backbone but integrates chemolithotrophs and taxa typically enriched under dry conditions or fluctuating moisture regimes.



- Cluster 3 (C and C+O t0) is the most distinctive cluster: it is dominated by *Candidatus Methylomirabilis oxyfera* (linked to nitrite and methane in anoxic microenvironments), together with *Thiobacillus thioparus* and *T. denitrificans*. Opportunistic/PGPR such as *Pseudomonas putida*, *P. fluorescens*, *P. umsongensis*, and *Peribacillus asahii* are also abundant. This profile is consistent with higher soil moisture and micro-anoxia, availability of reactive nitrogen, and labile carbon inputs (organic residues).

Supplementary material S4 (Table S4)

ITS dataset (microorganisms)	M t0 - M+O t0	C t0 - C+O t0	M+O t1	M t1	C+O t1	C t1	M+O t2	M t2	C+O t2	C t2	M+O t3	M t3	C+O t3	C t3	M+O t4	M t4	C+O t4	C t4	M+O t5	M t5	C+O t5	C t5
Actinomucor elegans	0,0000	0,0339	0,0000	0,0000	0,0000	0,0000	0,0125	0,0079	0,0000	0,0000	0,0000	0,0372	0,0000	0,0000	0,0136	0,0000	0,0000	0,0116	0,0000	0,0000	0,0000	0,0000
Actinomucor elegans var. kuwaitensis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0071	0,0000	0,0000
Ascomycota	0,0000	0,0000	0,0065	0,0000	0,0000	0,0112	0,0086	0,0102	0,0210	0,0268	0,0000	0,0000	0,0092	0,0000	0,0000	0,0066	0,0000	0,0000	0,0078	0,0000	0,0000	0,0000
Aureobasidium pullulans	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0109	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Basidiomycota	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0079	0,0113	0,0000	0,0000	0,0110	0,0220	0,0088	0,0000	0,0066	0,0071	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Bolbitiaceae	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0656	0,0000	0,0311	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Botryotrichum atrogriseum	0,0000	0,0000	0,0066	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Cladosporiaceae	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0217	0,0282	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0148	0,0000	0,0000	0,0000
Cladosporium halotolerans	0,0000	0,0000	0,0126	0,0000	0,0078	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Cladosporium sp.	0,0000	0,0000	0,0124	0,0000	0,0073	0,0000	0,0000	0,0000	0,0631	0,0948	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0566	0,0106	0,0000	0,0076
Conocybe ceracea	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0153	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Conocybe dunensis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,1175	0,0000	0,0445	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Conocybe fuscimarginata	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0170	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Conocybe incarnata	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0286	0,0000	0,0108	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Conocybe sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0507	0,0000	0,0132	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Coprinellus radians	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0191	0,0000	0,0000	0,0000	0,0087	0,0000	0,0000	0,0109	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Coprinellus verrucispermus	0,0268	0,0363	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Didymella sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0121	0,0089	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0063	0,0000	0,0000	0,0000
Epicoccum brasiliense	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0075	0,0126	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0070	0,0000	0,0000	0,0000
Epicoccum endophyticum	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0804	0,2196	0,1085	0,1347
Epicoccum nigrum	0,0000	0,0266	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,1761	0,2066	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,1100	0,0000	0,0000	0,0000
Epicoccum phragmospora	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0474	0,1265	0,0489	0,0459
Epicoccum sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0117	0,0097	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0785	0,1930	0,0992	0,1330

Hypocreaceae	0,0000	0,0000	0,0761	0,0744	0,0805	0,0749	0,0662	0,0636	0,0425	0,0327	0,0427	0,0495	0,0522	0,0817	0,0798	0,0811	0,0864	0,0730	0,0221	0,0000	0,0000	0,0000
Linnemannia amoeboidea	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0077	0,0105	0,0000	0,0000	0,0186	0,0347	0,0327	0,0082	0,0117	0,0000	0,0000	0,0000	0,0000	0,0100	0,0000	0,0000
Linnemannia elongata	0,0480	0,0387	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Linnemannia exigua	0,0332	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0078	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Linnemannia gamsii	0,0314	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0150	0,0144	0,0092	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Mortierella alpina	0,4866	0,2470	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Mortierella globalpina	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0114	0,0159	0,0116	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0086	0,0000	0,0000
Mortierella lapis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0085	0,0112	0,0000	0,0000	0,0313	0,0501	0,0418	0,0113	0,0142	0,0000	0,0000	0,0080	0,0000	0,0114	0,0000	0,0000
Mortierella sp.	0,0000	0,0000	0,0000	0,0066	0,0086	0,0080	0,0189	0,0339	0,0000	0,0000	0,0617	0,0715	0,0552	0,0116	0,0186	0,0155	0,0000	0,0245	0,0000	0,0000	0,0000	0,0000
Mortierellaceae	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0070	0,0000	0,0000	0,0000	0,0125	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Mrakia sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0479	0,0716	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0341	0,0000	0,0000	0,0000
Nectriaceae	0,0000	0,0000	0,0000	0,0071	0,0000	0,0157	0,0085	0,0078	0,0000	0,0000	0,0141	0,0182	0,0144	0,0000	0,0074	0,0000	0,0000	0,0154	0,0000	0,0000	0,0000	0,0000
Neodidymelliopsis sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0111	0,0074	0,0088
Nothophoma acaciae	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0076	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0108	0,0108	0,0065	0,0089
Nothophoma sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0085	0,0093	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0071	0,0000	0,0000	0,0000
Paraconiothyrium brasiliense	0,0637	0,0799	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Paramacroventuria ribis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0291	0,0459	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0345	0,0000	0,0000	0,0000
Paramacroventuria sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0266	0,0337	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0922	0,1902	0,0856	0,1033
Pezizales	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0107	0,0000	0,0000	0,0000	0,0080	0,0000	0,0118	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Plectosphaerella cucumerina	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0205	0,0118	0,0000	0,0000	0,0066	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Podila minutissima	0,0295	0,0823	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Preussia flanaganii	0,0222	0,0291	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Pseudeurotium bakeri	0,1754	0,2639	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0126	0,0103	0,0129
Pseudeurotium hygrophilum	0,0332	0,0363	0,0000	0,0000	0,0000	0,0123	0,0067	0,0075	0,0000	0,0000	0,0101	0,0098	0,0000	0,0000	0,0000	0,0000	0,0000	0,0109	0,0000	0,0000	0,0000	0,0000
Pseudeurotium sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0116	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Pyrenochaetopsis decipiens	0,0000	0,0363	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Rhizopus arrhizus	0,0240	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0115	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Saccharomyces cerevisiae	0,0000	0,0000	0,0000	0,0000	0,0137	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0069	0,0000	0,0000
Sporormiaceae	0,0000	0,0000	0,0000	0,0000	0,0000	0,0075	0,0000	0,0000	0,0000	0,0000	0,0076	0,0105	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0000	0,0000	0,0000	0,0000
Trichoderma asperellum	0,0000	0,0000	0,0107	0,0098	0,0087	0,0097	0,0089	0,0077	0,0000	0,0000	0,0000	0,0000	0,0000	0,0089	0,0089	0,0091	0,0097	0,0000	0,0000	0,0000	0,0000	0,0000
Trichoderma atroviride	0,0000	0,0000	0,0214	0,0234	0,0192	0,0326	0,0276	0,0309	0,0120	0,0000	0,0095	0,0122	0,0093	0,0218	0,0157	0,0333	0,0291	0,0154	0,0167	0,0139	0,0550	0,0515
Trichoderma austrokoningii	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0093	0,0153	0,0110	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Trichoderma caribbaeum	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0089	0,0221	0,0215
Trichoderma dorotheae	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0089	0,0095
Trichoderma gamsii	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0213	0,0084	0,0464	0,0262
Trichoderma koningiopsis	0,0000	0,0000	0,0261	0,0291	0,0254	0,0259	0,0194	0,0204	0,0163	0,0110	0,0157	0,0208	0,0189	0,0281	0,0245	0,0317	0,0353	0,0235	0,0078	0,0000	0,0000	0,0000
Trichoderma neokoningii	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0168	0,0256	0,0969	0,0821
Trichoderma ochroleucum	0,0000	0,0000	0,0774	0,0778	0,0751	0,0710	0,0657	0,0666	0,0445	0,0299	0,0345	0,0428	0,0499	0,0736	0,0771	0,0740	0,0748	0,0768	0,0214	0,0000	0,0000	0,0000
Trichoderma paucisporum	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0096	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Trichoderma scalesiae	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0138	0,0119
Trichoderma sp.	0,0000	0,0000	0,2303	0,2403	0,2397	0,2488	0,2469	0,2443	0,1301	0,1036	0,1038	0,1156	0,1555	0,2298	0,2093	0,2192	0,2413	0,2138	0,1248	0,1249	0,3905	0,3422
Trichoderma viride	0,0000	0,0000	0,1855	0,1863	0,1799	0,1706	0,1563	0,1625	0,1026	0,0685	0,0799	0,1022	0,1014	0,1745	0,1698	0,1720	0,1779	0,1648	0,0564	0,0000	0,0000	0,0000
Unclassified Fungi	0,0000	0,0000	0,3344	0,3451	0,3342	0,3002	0,2999	0,2968	0,2128	0,1764	0,2047	0,3127	0,2861	0,3395	0,3429	0,3330	0,3455	0,3522	0,1111	0,0000	0,0000	0,0000
Vacuiphoma ferulae	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0139	0,0222	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0141	0,0000	0,0000	0,0000
Westerdykella globosa	0,0259	0,0896	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Supplementary material S5 (Table S5)

16S dataset (microorganisms)	M t0 - M+O t0	C t0 - C+O t0	M+O t1	M t1	C+O t1	C t1	M+O t2	M t2	C+O t2	C t2	M+O t3	M t3	C+O t3	C t3	M+O t4	M t4	C+O t4	C t4	M+O t5	M t5	C+O t5	C t5
Aciditerrimonas ferrireducens JCM 15389	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0194	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0190	0,0000
Antarcticibacterium flavum	0,0000	0,0000	0,0593	0,0381	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Arenimicrobium luteum	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0313	0,0348	0,0000	0,0226	0,0279	0,0211	0,0000	0,0000	0,0385	0,0427	0,0224	0,0000	0,0297	0,0201	0,0207	0,0288
Bacillus badius	0,0000	0,0364	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Bacillus sp.	0,0000	0,0000	0,1048	0,0828	0,1001	0,1070	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Brevitalea deliciosa	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0327	0,0352	0,0000	0,0201	0,0269	0,0248	0,0000	0,0000	0,0362	0,0442	0,0196	0,0000	0,0268	0,0202	0,0190	0,0226
Brevitalea sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0241	0,0245	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Candidatus Methylomirabilis oxyfera	0,2343	0,1778	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Candidatus Udaeobacter copiosus	0,0590	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Chujaibacter soli	0,0000	0,0000	0,0000	0,0352	0,0322	0,0438	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Cognatiluteimonas lumbrici	0,0000	0,0000	0,0000	0,0000	0,0345	0,0456	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Cytobacillus spp.	0,0000	0,0000	0,0000	0,0000	0,0384	0,0327	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Gaiella occulta	0,1144	0,0485	0,0778	0,0840	0,0702	0,0584	0,0935	0,0944	0,1008	0,1080	0,1148	0,1222	0,1176	0,1331	0,0923	0,1074	0,0927	0,1550	0,1025	0,0816	0,0889	0,0874
Gemmatimonas aurantiaca T-27	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0342	0,0405	0,0328	0,0315	0,0341	0,0370	0,0271	0,0316	0,0371	0,0396	0,0195	0,0345	0,0254	0,0220	0,0287	0,0389
Gemmatimonas phototrophica	0,0000	0,0000	0,0000	0,0364	0,0000	0,0316	0,0342	0,0375	0,0380	0,0326	0,0297	0,0315	0,0282	0,0308	0,0390	0,0372	0,0193	0,0292	0,0214	0,0220	0,0230	0,0306
Gemmatirosa kalamazoonensis	0,0000	0,0000	0,0469	0,0397	0,0337	0,0000	0,0441	0,0482	0,0336	0,0382	0,0421	0,0484	0,0423	0,0531	0,0399	0,0442	0,0274	0,0402	0,0318	0,0337	0,0339	0,0463
Geoalkalibacter ferrihydriticus	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0163	0,0000	0,0000	0,0000
Limisphaera ngatamarikiensis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0269	0,0281	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Longimicrobium terrae	0,0000	0,0000	0,0604	0,0344	0,0340	0,0371	0,0389	0,0492	0,0483	0,0358	0,0366	0,0343	0,0390	0,0380	0,0320	0,0324	0,0207	0,0366	0,0271	0,0363	0,0339	0,0434
Luteitalea pratensis	0,1144	0,0727	0,1000	0,1018	0,0601	0,0523	0,1376	0,1422	0,0747	0,1104	0,1576	0,1398	0,1210	0,1201	0,1535	0,1477	0,0960	0,1121	0,1143	0,1099	0,0992	0,1223
Mesobacillus subterraneus	0,0000	0,0000	0,0000	0,0000	0,0281	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Neobacillus cucumis	0,0000	0,0000	0,0000	0,0000	0,0400	0,0401	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Neobacillus niacini	0,0000	0,0343	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Nitrosospira multiformis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0290	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Nocardioides sp.	0,0000	0,0000	0,0000	0,0000	0,0371	0,0283	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Nordella oligomobilis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,1070	0,0448	0,0448	0,0857	0,0821	0,0411
Nordella sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,2284	0,0931	0,0899	0,1567	0,1492	0,0768
Noviluteimonas caseinilytica	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0451	0,0443	0,0332	0,0348	0,0359	0,0410	0,0247	0,0329	0,0000	0,0000	0,0000	0,0241	0,0244	0,0294	0,0237	0,0000
Novilysobacter spongiicola	0,0000	0,0000	0,0000	0,0646	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Ohtaekwangia koreensis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0228	0,0000	0,0000	0,0197	0,0000	0,0000	0,0000	0,0000	0,0297	0,0000	0,0000	0,0000	0,0000	0,0000	0,0228	0,0238
Peribacillus asahii	0,0000	0,0949	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Povalibacter uvarum	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0252	0,0224	0,0237	0,0000	0,0260	0,0254	0,0169	0,0000	0,0171	0,0230	0,0261	0,0201
Priestia aryabhattai	0,0000	0,0000	0,0585	0,0447	0,0573	0,0641	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Priestia aryabhattai B8W22	0,0000	0,0000	0,0728	0,0621	0,0964	0,0929	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Priestia megaterium	0,0000	0,0000	0,0491	0,0439	0,0546	0,0578	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Priestia sp.	0,0000	0,0000	0,0000	0,0000	0,0384	0,0342	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Pseudomonas fluorescens	0,0000	0,0505	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Pseudomonas putida	0,0000	0,0889	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Pseudomonas sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0225	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Pseudomonas umsongensis	0,0000	0,0343	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Rhabdothermincola sediminis	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0228	0,0273	0,0312	0,0281	0,0231	0,0248	0,0290	0,0300	0,0000	0,0000	0,0203	0,0355	0,0231	0,0252	0,0259	0,0206
Roseisolibacter agri	0,0000	0,0000	0,0000	0,0389	0,0000	0,0329	0,0356	0,0359	0,0280	0,0273	0,0321	0,0345	0,0284	0,0340	0,0357	0,0357	0,0000	0,0281	0,0180	0,0000	0,0209	0,0279
Rossellomorea marisflavi	0,0000	0,0000	0,0000	0,0000	0,0000	0,0353	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Salinimicrobium soli	0,0000	0,0000	0,0643	0,0501	0,0286	0,0464	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Sphingomonas sediminicola	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0218	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Sphingomonas sp.	0,0000	0,0000	0,0522	0,0000	0,0351	0,0349	0,0351	0,0306	0,0421	0,0383	0,0293	0,0310	0,0341	0,0473	0,0329	0,0266	0,0000	0,0260	0,0182	0,0000	0,0000	0,0000
Stenotrophobacter terrae	0,0000	0,0424	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Tepidisphaera mucosa	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0299	0,0273	0,0239	0,0235	0,0328	0,0345	0,0243	0,0321	0,0413	0,0424	0,0000	0,0298	0,0000	0,0000	0,0000	0,0201
Thermoanaerobaculum aquaticum	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0186	0,0206	0,0000	0,0000
Thiobacillus denitrificans	0,0000	0,0444	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0270	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Thiobacillus sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0479	0,0302	0,0000	0,0000	0,0233	0,0351	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Thiobacillus thioparus	0,1568	0,1091	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,1594	0,1026	0,0000	0,0000	0,0804	0,0839	0,0000	0,0000	0,0211	0,0319	0,0000	0,0000	0,0218	0,0348
Thiobacillus thiophilus	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0704	0,0484	0,0000	0,0000	0,0318	0,0396	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Thiobacter subterraneus	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0313	0,0224	0,0000	0,0000	0,0266	0,0227	0,0286	0,0000	0,0283	0,0297	0,0177	0,0231	0,0252	0,0336	0,0243	0,0300
Tumebacillus ginsengisoli	0,0000	0,0000	0,0000	0,0000	0,0332	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Vicinamibacter silvestris	0,3210	0,1657	0,2539	0,2086	0,1480	0,1247	0,3093	0,3092	0,1572	0,2479	0,3041	0,2914	0,2738	0,2584	0,2866	0,2923	0,2534	0,2559	0,3066	0,2800	0,2366	0,2845
Vicinamibacter sp.	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0210	0,0000	0,0000	0,0214	0,0194	0,0228	0,0000	0,0000	0,0000	0,0177	0,0000	0,0188	0,0000	0,0000	0,0000
Vreelandella sp.	0,0000	0,0000	0,0000	0,0348	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

Chapter 4 – Predicting soil temperature from soil microbial community profiles with a Feed-Forward Artificial Neural Network

4.1 Introduction

Ecologically, soil temperature is a key driver of microbial activity, nutrient cycling, and community turnover (Wang et al. 2021; Qu et al. 2023). Because microbial communities can react to abiotic gradients in non-linear, interaction-rich ways (Yu et al. 2025), prediction models are a useful addition to simple correlations: it lets us quantify the temperature signal in the microbiome, while controlling for temporal structure and minimizing treatment confounding.

As demonstrated in Chapter 3, the treatments did not yield the expected treatment-specific trajectories in microbiome composition across time; instead, temporal dynamics predominated as the main source of community variation. Building on this result, Chapter 4 investigates the relationship between soil temperature and microbial community composition by pooling treatments within each sampling time and focusing on within-time variability across replicates. Specifically, an ANN was implemented to predict temperature from soil microbiome profiles, allowing for potentially nonlinear responses to environmental gradients without imposing treatment-specific effects, that were not supported by prior analyses.

Methodologically, the approach addresses the challenges of sparse, compositional, and collinear abundance data through careful pre-processing and supervised learning with nonlinear function approximation. Model design prioritizes generalization and interpretability, with rigorous cross-validation and complementary error metrics to quantify predictive accuracy. The objectives are threefold: (i) assess the feasibility of temperature prediction from microbial profiles; (ii) compare performance against simple baselines to establish added value; and (iii) probe model behaviour to identify taxa and community patterns most informative for prediction. By reframing microbial data as a proxy for latent environmental states, this work illustrates how AI, and in particular a feed-forward ANN, can deepen our understanding of soil microbiome–environment coupling and open avenues for microbially informed monitoring in agrifood systems.

4.2 Materials and methods

A complete end-to-end pipeline was implemented to (i) map absolute soil temperature to each sample, (ii) convert microbiome abundances to CLR space, (iii) build a supervised dataset, (iv) augment it synthetically, (v) train a small ANN regressor, (vi) evaluate it, and (vii) derive a simple "feature importance" proxy from the first hidden layer's weights.

For each sampling event, the average soil temperature over the 10 days preceding the sampling date was calculated to represent the most ecologically relevant window of thermal exposure. The use of antecedent 'climate windows' is well established in ecology, where biological responses often integrate environmental conditions over prior days to weeks (Cable et al. 2013). Temperature data were obtained

from Evja on-site soil sensors installed in the experimental field, with one probe monitoring the C/C+O plot and a second probe monitoring the M/M+O plot (Table 1 and Table 2).

Table 1. Mean soil temperature over the 10-day period preceding the sampling date for the M and M+O treatments

M/M+O	
Sampling time	**T°mean10d**
t0	6.5
t1	14.4
t2	18.5
t3	27.3
t4	24.0
t5	23.7

Table 2. Mean soil temperature over the 10-day period preceding the sampling date for the C and C+O treatments

C/C+O	
Sampling time	**T°mean10d**
t0	6.50
t1	13.50
t2	18.54
t3	27.51
t4	26.79
t5	20.31

Microbial community data from both 16S rRNA and ITS sequencing datasets were melted and matched to the corresponding 10-day mean temperature for each sample. Samples name encoded treatment, sampling time, and biological replicate (e.g., M t1 1, C+O t4 3).

These identifiers were parsed automatically to extract metadata on treatment (C, C+O, M, M+O) and time (t1–t5).

Soil temperature values (10-day means) were retrieved from two independent files corresponding to the C/C+O and M/M+O groups and matched to each sample according to its time point. Because microbial abundances are compositional, each sample vector was transformed to centered log-ratio (CLR) space (Aitchison 1982) used for regression-based analyses, ensuring that relationships among taxa are expressed as log-ratios rather than absolute abundances. A small pseudocount (e.g., 1×10^{-6}) was added to all abundances prior to transformation to avoid undefined logarithms caused by zero values, while minimally affecting the relative proportions among taxa. This numerical correction ensures the stability of the CLR transformation and preserves the overall structure of the compositional data. The use of small pseudocounts before CLR transformation is a common practice in compositional data analysis for microbiome studies (Gloor et al. 2017). Only samples that parsed unambiguously, had an assigned temperature, and produced finite CLR values were retained. The resulting CLR matrix served as predictors (X), and the corresponding absolute temperature as the target (y).

To increase training signal without imposing treatment structure, data augmentation was performed directly in CLR space. In this work, three augmentation strategies were employed to enhance model robustness and coverage of the feature space. First, zero-mean Gaussian perturbations with standard deviation σ = 0.005 were applied to each feature of observed samples, introducing minimal stochastic variability while leaving the target (soil temperature) unchanged (Bishop 1995). Second, within-treatment mix-up was performed by forming convex combinations of pairs of samples and their corresponding targets, with the mixing coefficient λ drawn from a Beta (0.3, 0.3) distribution; this yields synthetic observations that remain close to one parent sample while preserving continuity along treatment-specific manifolds (Zhang et al. 2017). Third, a SMOTE-like procedure was adopted whereby, for each sample, one of its three nearest neighbours (Euclidean distance) was selected and a new observation was generated by linear interpolation between the two; the target temperature was interpolated using the same proportion (Torgo et al. 2013). Collectively, these procedures introduce realistic, small-scale variation, promote smoother decision boundaries, and reduce sparsity in regions of the feature space while respecting treatment structure and biological plausibility. Original and synthetic samples were concatenated and randomly shuffled.

4.2.1 ANN Architecture

Temperature was predicted from CLR-transformed microbial profiles using a compact feed-forward multilayer perceptron designed for regression. The network comprised two hidden layers with 16 ReLU units each, a standard choice that supports efficient gradient flow and nonlinear function approximation (Nair & Hinton 2010; Goodfellow et al. 2016). A "ReLU unit" denotes a neuron applying the Rectified Linear Unit activation (Figure 1) to a linear pre-activation. For input x and parameters (w,b), the unit's output is:

$$\mathrm{ReLU}(\mathbf{w}^\top \mathbf{x} + b) \;=\; \max\!\left(0,\, \mathbf{w}^\top \mathbf{x} + b\right)$$

Figure 1. Rectified Linear Unit (ReLU). Neuron output defined, where w are the weights, x the input, and b the bias. ReLU zeroes negative activations and passes positive ones unchanged, introducing nonlinearity and stabilizing gradient flow.

The ReLU activation sets negative pre-activations to zero and leaves positive values unchanged, which promotes stable gradient propagation and enables the network to model nonlinear relationships at low computational cost. Having two hidden layers with 16 ReLU units each means that the first hidden layer produces 16 activated features from the input, and the second hidden layer takes those 16 values and applies the same operation again, allowing the model to compose simple nonlinearities into richer patterns. The final output layer then aggregates the 16 values from the second hidden layer into a single scalar prediction (temperature). In the present setting (CLR input with D=240 features), the first dense layer contains 240×16+16=3,856 parameters, the second 16×16+16=272, and the output layer

16×1+1=17. This configuration provides compact capacity, sufficient to capture signal in CLR-transformed microbial profiles, while helping control overfitting when combined with L2 regularization and dropout.

To mitigate overfitting, L2 weight decay ($\lambda = 0.001$) was applied to all dense layers, implementing Tikhonov regularization on the weights (Krogh & Hertz 1991; Bishop 2006), and dropout ($p = 0.15$) was used to decorrelate hidden activations (Srivastava et al. 2014). The output layer contained a single linear neuron suitable for continuous targets under a mean-squared-error objective (Bishop 2006). Input features, and the target during optimization, were z-standardized using training-set moments to stabilize optimization and improve conditioning (Hastie et al. 2009). Training proceeded for up to 100 epochs. An epoch denotes a full pass over the training set (excluding the portion held out for validation). In this run, the training partition contained X_train = 154 samples; with validation_split = 0.2, approximately 123 samples per epoch were effectively used to update the weights. Given a mini-batch size = 16, the number of weight updates per epoch is 123/16 = 8 updates per epoch.

The network was optimized with Adam (learning rate 0.001), a first-order adaptive method that combines momentum and per-parameter scaling (Kingma and Ba 2014), minimizing MSE, using a random 90/10 train–test split, an internal 20% validation split. Early stopping on validation loss (patience = 50, best-weights restore) and learning-rate reduction on plateau (factor = 0.5, patience = 10; floor = 1×10^{-6}) were employed to prevent overfitting and aid convergence once progress slowed (Goodfellow et al. 2016). This setup intentionally pooled treatments and accepted potential leakage from random splitting, in line with the study aim of exploring nonlinear relationships between temperature and community composition.

The resulting architecture balances model capacity and regularization, providing a lightweight, robust regressor for mapping CLR features to absolute soil temperature.

Predictive performance and diagnostics were assessed on the held-out test set using several metrics (Rohart et al. 2017):

- Coefficient of determination (R^2) (Figure 2) on the real °C scale, quantifying the proportion of variance in observed temperatures explained by model predictions (higher is better)

$$R^2 = 1 - \frac{\sum_n (y_n - \hat{y}_n)^2}{\sum_n (y_n - \bar{y})^2}$$

Figure 2. R^2 (goodness of fit). Indicates how much prediction error is reduced compared with using the sample mean; higher values mean better fit.

- Mean absolute error (MAE) (Figure 3) in °C, reporting the average absolute deviation between predicted and observed temperatures (lower is better)

$$\mathrm{MAE} = \frac{1}{N}\sum_{n=1}^{N} |y_n - \hat{y}_n|$$

Figure 3. Mean Absolute Error (MAE). Average absolute difference between observed and predicted values.

- Root mean square error (RMSE) (Figure 4) in °C, emphasizing larger deviations through squaring before averaging (lower is better).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n-1}^{N}(y_n - \hat{y}_n)^2}$$

Figure 4. Root Mean Squared Error (RMSE) Typical prediction error in the original units, with greater weight on large deviations.

Model behaviour was further examined through diagnostic visualizations:

- Training/validation loss and MAE curves across epochs, to evaluate optimization progress, potential overfitting (divergence), and convergence under early stopping.
- Predicted vs. observed temperature scatterplot with a 1:1 reference line, to assess calibration, dispersion around the identity, and systematic bias across the temperature range.
- Residual distribution (observed − predicted), to inspect error symmetry, central tendency (mean near zero), and the presence of heavy tails or outliers.

To provide an interpretable signal of contributing taxa, the absolute input-to-first-hidden-layer weights were averaged per feature, and taxa were ranked by mean absolute weight; the top contributors were reported with the caution that deeper nonlinear interactions are not captured by this proxy.

4.2.2 Software and Data Management

All analyses described in this chapter were conducted within a reproducible computational environment to ensure consistency, transparency, and traceability of results. The entire workflow was developed in Python 3.11 (Anaconda distribution) and executed through Jupyter Notebooks, allowing the integration of code, results, and documentation within a single, version-controlled environment.

4.2.3 Computational environment

The analyses were carried out on a Windows 10 workstation equipped with 32 GB RAM and an Intel® Core™ i7 processor. The following Python libraries were employed:

- Data handling and preprocessing*: pandas (I/O, tabular wrangling), numpy (vectorized computation), warnings (non-critical warning suppression*
- ML and related analyses*: tensorflow.keras (feed-forward ANN, training callbacks), scikit-learn (train/test split, standardization via StandardScaler, nearest neighbours for SMOTE-like interpolation, and evaluation metrics).*
- Visualization: *matplotlib (training curves, predicted-vs-observed plots, residual distributions, and feature-importance bar charts).*

All packages were managed via Anaconda environment to guarantee library compatibility and reproducibility.

4.2.4 Data organization and preprocessing

The primary input data consisted of Excel (.xlsx) tables derived from metagenomic sequencing outputs (16S and ITS), where:

- Rows represented identified taxa (genera or species level).
- Columns corresponded to individual samples (biological replicates defined by treatment and sampling time).
- Cells contained normalized read counts or relative abundances.

Each dataset underwent initial preprocessing steps including:

- Removal of low-abundance and unclassified taxa.
- Normalization via Total Sum Scaling (TSS).
- Log transformation (when required) to stabilize variance prior to statistical testing.
- Merging of metadata files containing treatment, sampling time, temperature, and yield information.

These preprocessing steps were standardized through a custom Python function to minimize human error and ensure identical treatment across datasets.

4.2.5 File management and reproducibility

All scripts, raw data, and processed outputs were stored in a structured directory system. This structure allowed systematic tracking of data flow from input to final output and facilitated integration into the overall AI-based analysis pipeline.

4.2.6 Data validation and backup

Intermediate results and figures were automatically exported as .xlsx and .png files for verification. Data integrity was verified by random cross-checks between raw and processed files. Regular backups were performed on a secure drive, ensuring data safety and long-term accessibility.

4.2.7 Documentation and code availability

Each Jupyter Notebook was annotated with detailed markdown cells describing input files, parameter settings, and statistical rationale. These annotated notebooks form a complete computational record of the analyses and can be made available upon reasonable request.

This computational and data management strategy ensured robustness and reproducibility of all analyses presented in Chapter 4.

4.3 Results

The microbiome matrix comprised 240 taxa profiled across 61 real samples. Absolute soil temperature associated with each sample spanned 6.50–27.51 °C; within the training set, the target distribution had a mean of 21.01 °C and SD = 5.70 °C. Temperature trajectories were nearly identical across treatments

up to t2, but diverged at t3–t5, with C/C+O exhibiting higher temperatures at t3–t4 and lower at t5 relative to M/M+O (e.g., t3: 27.51 °C in C/C+O vs 27.25 °C in M/M+O; t4: 26.79 °C vs 23.96 °C; t5: 20.31 °C vs 23.69 °C).

Given the small number of samples relative to the number of features (small n, high p), data augmentation expanded the dataset from 61 to 172 instances (+61 Gaussian-noise, +20 mix-up, +30 SMOTE-like). A stratified split yielded 154 training and 18 test samples. A compact feed-forward network was trained (240→16→16→1, dropout on both hidden layers; ~4.1k parameters) with Adam and learning-rate scheduling. Validation error decreased steadily in the first third of training and then plateaued. Whenever no improvement was observed for several epochs, a ReduceLROnPlateau scheduler decreased the learning rate (triggers at epochs 36, 46, 56, 66, and 100), enabling finer weight updates. At the end of training, model weights were restored to those from the best validation epoch (epoch 90), which we used for all reported test metrics.

During training, both loss (MSE) and MAE (Figure 5 and 6) decline sharply in the first epochs and then flatten for training and validation in tandem, a pattern consistent with stable learning rather than overfitting. In the scatterplot of predicted versus observed temperatures (Figure 7), the points lie close to the 1:1 line and the coefficient of determination is high ($R^2 = 0.960$), indicating that the model reconstructs almost all of the variability in measured temperatures from microbial composition. The histogram of residuals (Figure 8) shows errors tightly centered around zero with a small average bias: predictions are, on average, about half a degree higher than the true values. Most deviations remain within roughly ±1 °C, suggesting accurate and reliable predictions across the temperature range sampled.

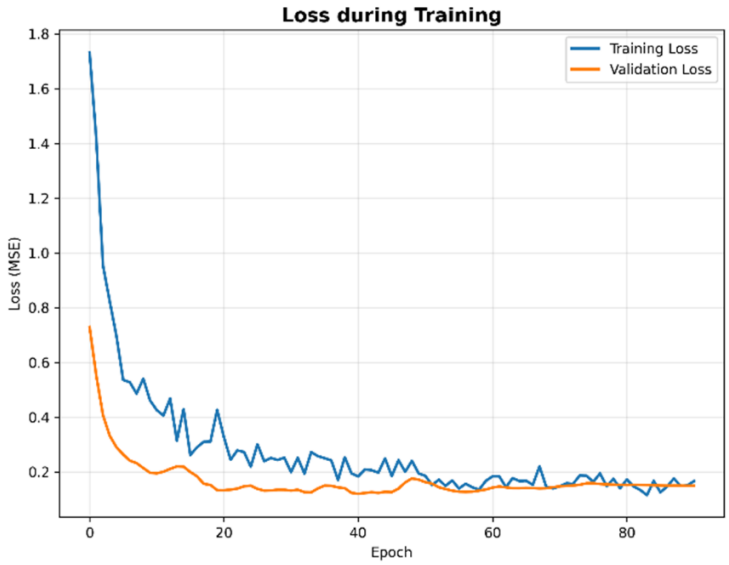


Figure 5. Loss (MSE) during training: epoch-wise training (blue) and validation (orange) losses illustrating stabilization after ~30–40 epochs and no divergence between curves.

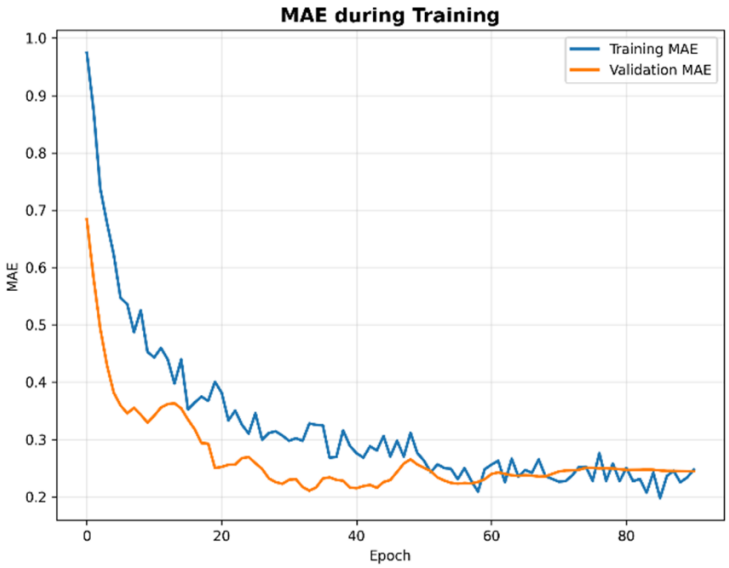


Figure 6. MAE during training: training (blue) and validation (orange) MAE decrease sharply and stabilize after ~40–60 epochs (plateau), with no sustained gap, suggesting the model has learned a consistent mapping without divergence between sets.

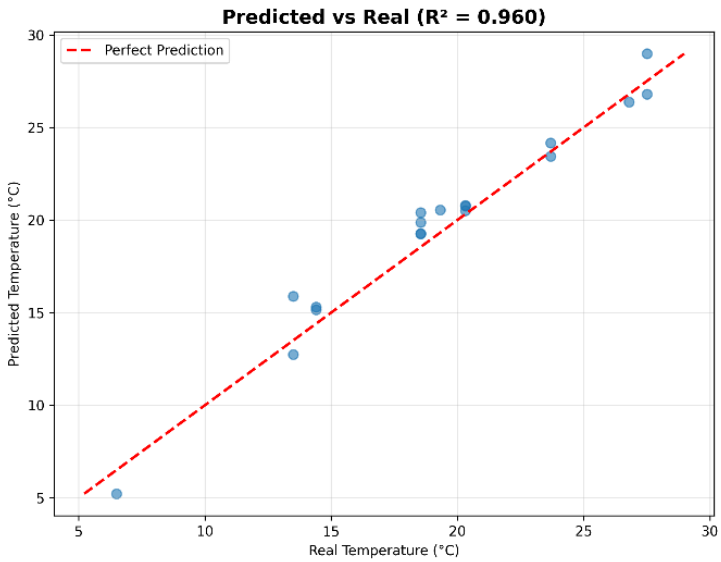


Figure 7. Scatterplot of predicted versus observed temperatures: predicted temperatures align tightly with observations against the 1:1 reference, with high explained variance ($R^2 = 0.960$).

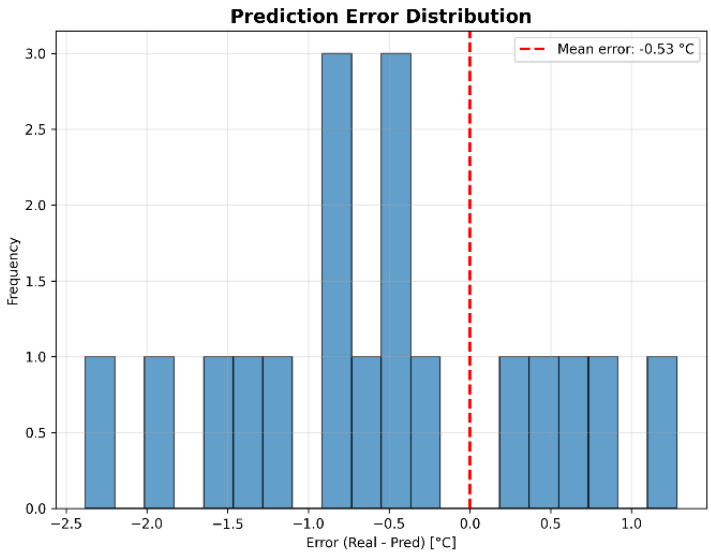


Figure 8. Histogram of residuals errors: errors cluster near zero but are centered at −0.53 °C, showing modest bias toward higher predicted temperatures.

Figure 9 complements the analysis by revealing which microbial taxa contribute most to those accurate predictions. It ranks the top twenty taxa by their overall importance to the model, effectively indicating which groups the network leans on most when distinguishing warmer from cooler samples. The bar lengths should be interpreted as predictive contribution rather than direction: the plot does not say whether a taxon increases or decreases with temperature, only how informative it is for the model. Read together, the results show that a compact subset of taxa encodes much of the temperature information and that the network can translate this signal into precise, low-error estimates of soil temperature.

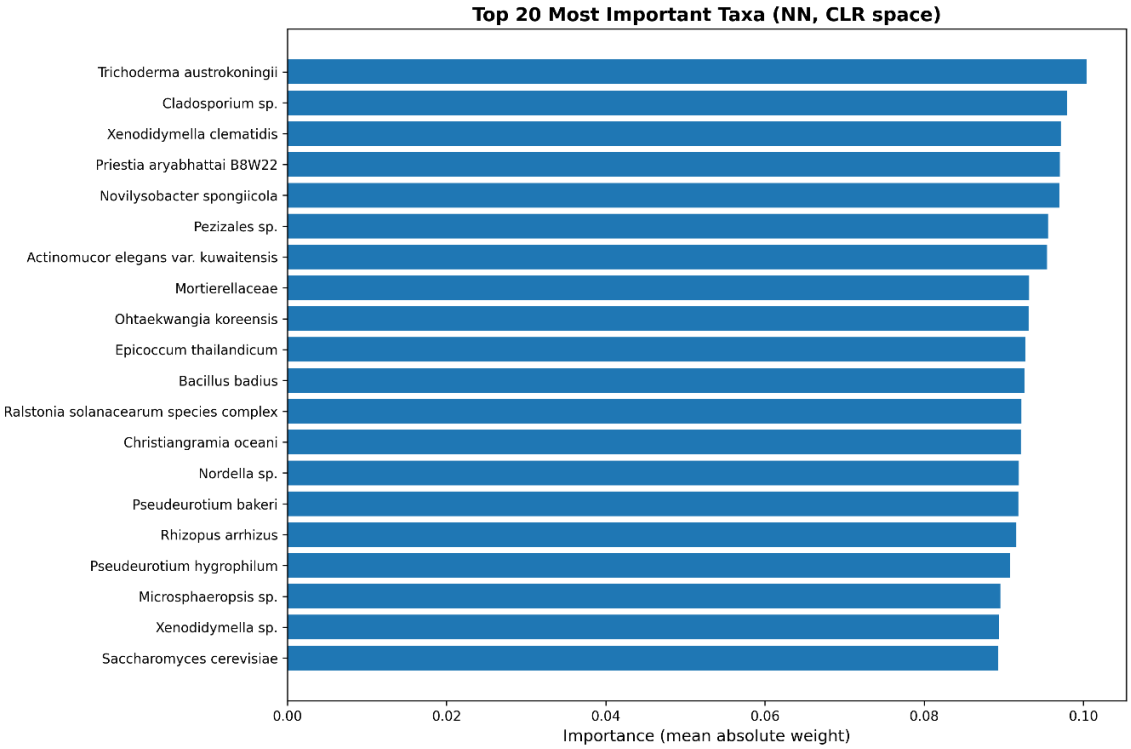


Figure 9. Bar plot of the 20 taxa with highest mean absolute weights in CLR space, indicating strongest contributions to the model.

4.4 Discussions

This study shows that soil temperature leaves a measurable "imprint" on microbiome composition that the compact feed-forward neural network implemented can decode with high accuracy. Using CLR-transformed abundances of 240 taxa, the model reconstructed nearly all observed temperature variability ($R^2 = 0.960$) with small average error (≈0.5 °C overestimation) and narrow residual dispersion (most errors within ±1 °C). These results indicate that microbial community structure contains a robust thermal signal across the sampled range (6.5–27.5 °C).

4.4.1 Methodological considerations

Working in a small-n, high-p regime required regularization and careful training dynamics. The architecture (240→16→16→1 with dropout and learning-rate scheduling) favoured a parsimonious hypothesis space that converged steadily without overfitting, as indicated by parallel training/validation curves and a flat late-epoch regime. CLR transformation was essential to respect compositional constraints and reduce spurious correlations. Data augmentation (Gaussian perturbations, within-treatment mix-up, SMOTE-like interpolation) increased sample diversity and improved the estimator's bias–variance trade-off, enabling stable learning from 61 to 172 instances. Although augmentations cannot create new ecological information, they help the optimizer explore smoother decision surfaces around observed points, which is crucial in this case.

4.4.2 Management context and temporal structure

Temperature trajectories were largely shared across treatments up to t2 and diverged at later time points (t3–t5), with C/C+O warmer at t3–t4 and cooler at t5 relative to M/M+O. The model's accuracy despite this divergence suggests that the thermal imprint is not treatment-idiosyncratic but is instead captured by community features that generalize across management regimes. This agrees with previous evidence in the dataset that time was the dominant driver of community variation, whereas treatment effects were weaker or more context dependent. In other words, the learned mapping appears to track a cross-treatment thermal gradient encoded in the microbiome rather than treatment-specific signals.

4.4.3 Ecological interpretation

The ability to infer temperature from composition suggests that community turnover co-varies with thermal conditions in a consistent way, likely reflecting temperature-sensitive metabolic rates, resource processing, and biotic interactions.

Among the twenty most informative features, several taxa align with saprotrophic or soil-adapted functional groups that plausibly encode recent thermal exposure. *Trichoderma austrokoningii* represents filamentous fungi with aggressive saprotrophy and mycoparasitism, supported by rich extracellular enzyme repertoires (Samuels et al. 2006). *Cladosporium* spp. are ubiquitous, mostly saprotrophic fungi that track litter turnover and near-surface microclimate; aerobiological records consistently show higher airborne *Cladosporium* loads under warmer conditions, with temperature emerging as a primary positive driver across multi-year datasets (Kasprzyk et al. 2016). *Xenodidymella* spp. (Didymellaceae) are commonly endophytic/saprotrophic on plant residues (Aveskamp et al. 2010), consistent with temperature-modulated detrital processing (Davidson and Janssens 2006). The bacterial contributors *Priestia aryabhattai* (formerly *Bacillus aryabhattai*) and *Bacillus badius* are spore-forming soil heterotrophs associated with polymer degradation and plant-associated traits (Yadav and Chandra 2016; Magome et al. 2024), whose metabolic rates typically scale with moderate warming (Nottingham et al. 2019). *Novilysobacter spongiicola* (Lysobacter-like) is linked to predatory/exoenzyme-rich strategies, indicative of active organic-matter turnover (Romanenko et al. 2008). Mucoralean fungi, such as *Actinomucor elegans* var. *kuwaitensis* and *Rhizopus arrhizus*, are fast-growing, hydrolytic decomposers that respond to thermo-hydric pulses (Wang et al. 1974; Han et al. 2003). Across the remaining contributors, several saprotrophic, residue-associated fungi, such as *Mortierellaceae, Epicoccum thailandicum*, *Pseudoeurotium bakeri*, *Pseudoeurotium hygrophilum*, and *Microsphaeropsis* sp., mark phases of litter colonization whose sporulation and abundance typically tracks thermo-osmotic conditions at the soil–litter interface (Ramos et al. 2021).

The prominence of saprotrophic and soil-adapted groups among the most informative features is ecologically coherent. These taxa cluster in functionally coherent groups, filamentous saprotrophs, exoenzyme-rich decomposers, and spore-forming soil bacteria, that are expected to co-vary with temperature via direct and indirect pathways (Tan et al. 2021). Direct effects include enzyme kinetics, membrane fluidity, and growth rates (Hassan et al. 2020; Davidson and Janssens 2006); indirect effects

arise when temperature alters substrate, redox dynamics, or plant inputs, which then feedback on microbial competition and cooperation (Gao et al. 2024). As a result, the same thermal increment can have different compositional consequences depending on moisture, oxygen, and recent disturbance.

Finally, because the features are compositional (CLR space), informativeness concentrates in groups that move in coordinated ways along thermal gradients. This does not guarantee that each taxon increases monotonically with temperature; some may show threshold-like or unimodal responses, and correlated clades can share importance due to co-variation.

4.5 Conclusions

A reliable microbiome-to-temperature mapping opens two applied avenues. First, it provides a complementary, biology-aware indicator of field thermal conditions that can be integrated with sensor networks, filling gaps and enabling retrospective inference when sensor data are missing. In practice, model predictions obtained from routine soil sampling can be assimilated alongside sensor data to fill temporal gaps and to backfill periods prior to sensor installation. Second, the same features that encode temperature may mediate plant–soil processes, such as decomposition or nutrient cycling, relevant to crop performance; linking the model's informative taxa to functional outcomes could support decision tools that combine environmental exposure and microbiome state to anticipate agronomic responses.

The same framework could enable spatial upscaling. Composite or georeferenced cores collected across a plot can be translated into temperature fields, yielding fine-grain maps of recent thermal exposure without saturating the field with hardware. These maps can be aligned to management zones and experimental blocks, facilitating comparisons among treatments when microclimate differs subtly across the site.

Because the model is lightweight, predictions can be produced rapidly, for example in a Python pipeline after sequencing outputs are processed, making it feasible to iterate as new samples arrive during the season.

Agronomically, the taxa that carry the temperature signal are not merely markers: many are functionally tied to decomposition and nutrient cycling. If increases in the abundance of saprotrophic or filamentous groups co-occur with warmer conditions, one may anticipate faster residue turnover and earlier nitrogen mineralization, with implications for fertilizer timing and cover-crop termination. Conversely, signals consistent with cooler exposure and slowed saprotrophy can motivate the retention of residues or delayed planting to avoid transient nutrient immobilization. In systems prone to heat or cold stress, the microbiome-derived temperature proxy can be combined with plant thermal thresholds to flag windows of elevated risk, particularly where canopy or soil sensors are sparse or unreliable.

For monitoring programs, standardized sampling (depth, timing within the day, soil moisture) and turnaround time from DNA to prediction are equally important: consistent protocols reduce variance unrelated to temperature, while a predictable pipeline enables timely use, ideally integrating with existing Python analysis scripts that process sequencing tables, apply CLR, load the trained model, and emit both temperature estimates and uncertainty metrics.

Finally, the approach is best viewed as a complement, not a replacement, to physical sensing. Where sensors are dense and clean, the microbiome proxy adds redundancy and ecological context; where sensors are sparse or intermittent, it provides a principled way to reconstruct recent exposure and to connect that exposure to microbially mediated processes that matter for crop performance. In both cases, translating model-informative taxa into functional groupings (e.g., saprotrophs vs. pathogens vs. mutualists) strengthens agronomic interpretation and helps close the loop from environmental exposure, to microbiome state, to actionable management decisions.

In sum, a small, regularized neural network trained on CLR-transformed microbiome profiles accurately recovers soil temperature and identifies a compact set of informative taxa that likely reflect temperature-sensitive ecological processes. The findings support the view that microbial communities carry a consistent thermal imprint and motivate integrative, function-aware modelling to translate that imprint into actionable indicators for soil and crop management.

4.6 References

Aitchison J (1982) The statistical analysis of compositional data. J R Stat Soc B 44: 139–177

Aveskamp MM, de Gruyter J, Woudenberg JHC, Verkley GJM, Crous PW (2010) Highlights of the Didymellaceae: a polyphasic approach to characterise Phoma and related pleosporalean genera. Stud Mycol 65: 1–60

Bishop CM (1995) Training with noise is equivalent to Tikhonov regularization. Neural Comput 7: 108–116

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Cable JM, Ogle K, Barron-Gafford GA, Bentley LP, Cable WL, Scott RL, Williams DG, Huxman TE (2013) Antecedent conditions influence soil respiration differences in shrub and grass patches. Ecosystems 16: 1230–1247

Davidson EA, Janssens IA (2006) Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. Nature 440: 165–173

Gao Y, Wang H, Yang F, Dai X, Meng S, Hu M, Fu X (2024) Relationships between root exudation and root morphological and architectural traits vary with growing season. Tree Physiol 44: tpad118

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome datasets are compositional: and this is not optional. Front Microbiol 8: 2224

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge, MA

Han BZ, Ma Y, Rombouts FM, Nout MR (2003) Effects of temperature and relative humidity on growth and enzyme production by Actinomucor elegans and Rhizopus oligosporus during sufu pehtze preparation. Food Chem 81: 27–34

Hassan N, Anesio AM, Rafiq M, Holtvoeth J, Bull I, Haleem A, Hasan F (2020) Temperature driven membrane lipid adaptation in glacial psychrophilic bacteria. Front Microbiol 11: 824

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. 2nd edn. Springer, New York

Kasprzyk I, Kaszewski BM, Weryszko-Chmielewska E, Nowak M, Sulborska A, Kaczmarek J, Jedryczka M (2016) Warm and dry weather accelerates and elongates Cladosporium spore seasons in Poland. Aerobiologia 32: 109–126

Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980

Krogh A, Hertz J (1991) A simple weight decay can improve generalization. In: Advances in neural information processing systems 4. Morgan Kaufmann, San Mateo, CA, pp 950–957

Magome TG, Ochai SO, Hassim A, Bezuidenhout CC, Van Heerden H, Lekota KE (2024) A genome-based investigation of the Priestia species isolated from anthrax endemic regions in Kruger National Park. Infect Genet Evol 123: 105649

Nair V, Hinton GE (2010) Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). Omnipress, Madison, WI, pp 807–814

Nottingham AT, Bååth E, Reischke S, Salinas N, Meir P (2019) Adaptation of soil microbial growth to temperature: using a tropical elevation gradient to predict future changes. Glob Change Biol 25: 827–838

Qu R, Liu G, Yue M, Wang G, Peng C, Wang K, Gao X (2023) Soil temperature, microbial biomass and enzyme activity are the critical factors affecting soil respiration in different soil layers in Ziwuling Mountains, China. Front Microbiol 14: 1105723

Ramos SM, Graça MA, Ferreira V (2021) A comparison of decomposition rates and biological colonization of leaf litter from tropical and temperate origins. Aquat Ecol 55: 925–940

Rohart F, Gautier B, Singh A, Lê Cao KA (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol 13: e1005752

Romanenko LA, Uchino M, Tanaka N, Frolova GM, Mikhailov VV (2008) Lysobacter spongiicola sp. nov., isolated from a deep-sea sponge. Int J Syst Evol Microbiol 58: 370–374

Samuels GJ, Dodd SL, Lu BS, Petrini O, Schroers HJ, Druzhinina IS (2006) The Trichoderma koningii aggregate species. Stud Mycol 56: 67–133

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15: 1929–1958

Tan B, Yin R, Zhang J, Xu Z, Liu Y, He S, Peng C (2021) Temperature and moisture modulate the contribution of soil fauna to litter decomposition via different pathways. Ecosystems 24: 1142–1156

Torgo L, Ribeiro RP, Pfahringer B, Branco P (2013) SMOTE for regression. In: Portuguese Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, pp 378–389

Wang C, Morrissey EM, Mau RL, Hayer M, Piñeiro J, Mack MC, Marks JC, Bell SL, Miller SN, Schwartz E, Dijkstra P, Koch BJ, Stone BW, Purcell AM, Blazewicz SJ, Hofmockel KS, Pett-Ridge J, Hungate BA (2021) The temperature sensitivity of soil: microbial biodiversity, growth, and carbon mineralization. ISME J 15: 2738–2747

Wang HL, Vespa JB, Hesseltine CW (1974) Acid protease production by fungi used in soybean food fermentation. Appl Microbiol 27: 906–911

Yadav B, Chandra K (2016) Isolation, purification, and characterization of xylanase produced by three species of Bacillus under submerged fermentation conditions. J Appl Nat Sci 8: 1210–1217

Yu K, He L, Niu S, Wang J, García-Palacios P, Dacal M, Averill C, Georgiou K, Ye J, Mo F, Yang L, Crowther TW (2025) Nonlinear microbial thermal response and its implications for abrupt soil organic carbon responses to warming. Nat Commun 16: 2763

Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2017) mixup: beyond empirical risk minimization. arXiv preprint arXiv:1710.09412

Chapter 5 – Conclusion and future perspectives

5.1 Original contributions of the PhD

The thesis delivers an end-to-end, reproducible framework that integrates compositional preprocessing and supervised models to quantify environmental imprints on microbiomes. The workflow explicitly handles small-n/high-p (few samples and many predictors), settings via regularization, meticulous data splits, and cautious evaluation to limit overfitting. A modular Python pipeline standardizes inputs/outputs (multi-sheet Excel, publication-ready figures), encapsulates model training and inference, and enables rapid iteration as new samples arrive. The design emphasizes transparency, portability, and ease of integration with routine sequencing outputs.

5.2 Overview and synthesis of key findings

This thesis investigated how soil microbial communities change through time and agricultural management, and whether those changes encode an actionable environmental signal, specifically soil temperature, that can be decoded with machine learning. By combining 16S/ITS community profiling, unsupervised learning, and a compact feed-forward neural network trained on CLR-transformed abundances, the work aimed to disentangle temporal versus treatment effects on community assembly, characterize functional succession across the crop cycles, and test whether community composition contains a reliable "thermal imprint" that is predictable at high resolution.

Across analyses, time, and the establishment of *Valerianella locusta*, emerged as the primary driver of microbial community structure, with treatment effects weaker and context-dependent. Communities at t0 reflected bulk-soil conditions dominated by decomposers and metabolically specialized bacteria; from t1 onward, the rhizosphere effect reshaped community composition, progressively favouring antagonistic and plant-associated fungi and oligotrophic, stress-tolerant bacterial groups. These dynamics reflect a shift from decomposer-rich bulk soils to rhizosphere-structured assemblages that stabilize along the crop cycle. Fungal assemblages showed higher plasticity and stronger sensitivity to management inputs (with antagonistic taxa enriched under inoculation and/or organic matter), whereas bacterial communities were more resilient, converging toward a functionally conserved core across treatments. Unsupervised learning (K-means, t-SNE) exposed these successional tracks and clarified subtle divergences not apparent from raw composition alone.

An ANN reconstructed nearly all observed temperature variability from CLR-transformed community profiles ($R^2 \approx 0.96$, small average overestimation, most residuals within ±1 °C), indicating that microbial composition carries a consistent, generalizable thermal signal over 6.5–27.5 °C. The model stays accurate even when treatments diverge late in the season, implying the predictive signal is shared across treatments, not driven by any one of them. Among the most informative features, saprotrophic and soil-adapted taxa (e.g., filamentous decomposers, exoenzyme-rich groups, spore-forming soil bacteria) predominated, ecologically coherent with temperature-sensitive turnover.

Taken together, these results suggest a practical role for microbiome profiling in monitoring and decision support. Community composition carries a biologically grounded record of recent thermal exposure, offering a complementary signal to physical sensors. Where sensor networks are sparse or discontinuous, microbiome-derived estimates can help backfill gaps and provide ecological context for interpreting temperature fluctuations, turning routine soil sampling into an additional layer of environmental surveillance.

The functional identity of the most informative taxa also speaks directly to management timing. Taxa linked to saprotrophic activity and filamentous growth tend to rise under warmer conditions, a pattern consistent with faster residue turnover and earlier nitrogen mineralization. In such cases, managers may anticipate advancing fertilizer applications or adjusting cover-crop termination to align nutrient release with crop demand. Conversely, profiles indicative of cooler exposure and subdued saprotrophy may signal transient immobilization risks, arguing for more conservative fertilization or delayed planting. In both directions, the microbiome serves as an integrative readout that bridges recent exposure with process-level consequences.

A further implication concerns system resilience. The observed convergence of bacterial communities toward a conserved core suggests functional redundancy that may buffer short-term management changes, whereas fungi display greater plasticity and responsiveness to inputs. This asymmetry invites a differentiated strategy: use bacterial stability as a baseline indicator of system robustness, while leveraging fungal responsiveness, particularly antagonistic or decomposer, for targeted interventions, such as organic amendments or biocontrol, always interpreted within the temporal trajectory of the crop cycle.

5.3 Future perspectives

The present findings originate from a single site and cropping system, which constrains their generality. Establishing transportability to other soils, climates, and management regimes remains a priority. Moreover, the limited set of physico-chemical and microclimatic covariates restricts attribution among correlated drivers, most notably temperature, moisture, substrate quality, and redox conditions. While CLR transformation mitigates closure, compositional dependencies and inter-taxon correlations can still bias feature importance and complicate interpretation. Finally, data-augmentation strategies (noise, mix-up, SMOTE-like interpolation) improve optimization in data-sparse settings but do not introduce new ecological information; their benefits should therefore be weighed against potential artifacts when extrapolating beyond the observed domain.

Advancing interpretability is central. In compositional data, correlated taxa share variance, so importance can cluster within functional groups and cannot be taken as evidence of monotonic, one-to-one relationships with temperature. Follow-up analyses, such as permutation importance, SHAP values, and partial-dependence or accumulated local effects profiles, should be used to establish directionality (increasing, decreasing, threshold-like), distinguish additive from interaction-driven effects, and test

robustness to correlation structure. The concentration of importance in ecologically coherent groups is encouraging, but mechanistic claims require these additional diagnostics.

On the data side, future work should enrich community profiles with environmental covariates, including pH, EC, soil moisture, organic C, total N, available P/K, texture, and paired soil/air microclimate. Derived features can capture ecologically relevant exposure windows. Within this expanded data space, multi-task learning that jointly predicts temperature and other abiotic states may promote shared representations and improve generalization.

To establish external validity, multi-site, multi-season designs with independent replication are recommended, accompanied by pre-registered analysis plans and strict out-of-distribution evaluation on held-out external test sets. Where feasible, pairing georeferenced soil cores with dense sensor grids would enable rigorous benchmarking of microbiome-derived thermal maps against physical measurements and facilitate spatial upscaling.

In sum, the path forward combines data enrichment, external validation, uncertainty-aware and interpretable modelling. Together, these steps will convert the present framework from a robust proof of concept into a reliable, transferable tool for sustainable soil and crop management.

v. Abbreviations

AI – Artificial Intelligence

ANN – Artificial Neural Network

CNN – Convolutional Neural Network

CLR – Centered Log-Ratio

DL – Deep Learning

DSS – Decision Support System

GBM – Gradient Boosting Machine(s)

MAE – Mean Absolute Error

ML – Machine Learning

MSE – Mean Squared Error

NGS – Next Generation Sequencing

ONT – Oxford Nanopore Technology

OP – Organization of Producers

PCA – Principal Component Analysis

PCR – Polymerase Chain Reaction

RF – Random Forest

RMSE – Root Mean Square Error

RNN – Recurrent Neural Network

SHAP – SHapley Additive exPlanations

SMOTE – Synthetic Minority Over-sampling Technique

SMRT – Single-Molecule Real-Time

SVM – Support Vector Machines

t-SNE – t-distributed Stochastic Neighbor Embedding

TSS – Total Sum Scaling

UMAP – Uniform Manifold Approximation and Projection