OXFORD UNIVERSITY PRESS | Briefings In Bioinformatics

MINERVA – Microbiome Network Research and Visualization Atlas: A Scalable Knowledge Graph for Mapping Microbiome-Disease Associations

Journal:	*Briefings in Bioinformatics*
Manuscript ID	BIB-25-0513
Manuscript Type:	Problem solving protocol
Date Submitted by the Author:	18-Mar-2025
Complete List of Authors:	Langarica, Saul; Pontificia Universidad Católica de Chile Escuela de Ingeniería, Electrical Engineering; Massachusetts General Hospital, Radiology; Harvard Medical School Kim, Young-Tak; Massachusetts General Hospital, Radiology; Harvard Medical School Alkhadrawi, Adham; Massachusetts General Hospital, Radiology; Harvard Medical School Kim, Jung Bin; Korea University College of Medicine, Department of Neurology Do, Synho; Harvard Medical School; Korea University; Massachusetts General Hospital, Radiology; Harvard University, Kemper Institute
Keywords:	Microbiome, Large Language Models, Knowledge Graph, Ontology, Gut Microbes

MINERVA – Microbiome Network Research and Visualization Atlas:

A Scalable Knowledge Graph for Mapping Microbiome-Disease Associations

Authors: Saul Langarica[1,2], Young-Tak Kim[1], Adham Alkhadrawi[1], Jung Bin Kim[3], Synho Do[1,4,5*]

Affiliations:

[1] Laboratory of Medical Imaging and Computation, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, 125 Nashua Street, Boston, Massachusetts, USA.

[2] Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, Chile.

[3] Department of Neurology, Korea University Anam Hospital, Korea University College of Medicine, 73, Goryeodae-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea.

[4] KU-KIST Graduate School of Converging Science and Technology, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea.

[5] Kempner Institute, Harvard University, 150 Western Ave, Boston, Massachusetts, USA.

*Corresponding author:

Email: sdo@mgh.harved.edu

Phone: 339-222-4409

Abstract:

Bacterial pathogens contribute significantly to the global burden of disease. Understanding their complex interactions with human health is essential for developing new diagnostic, preventative, and therapeutic strategies. While recent breakthroughs have revolutionized our understanding of these relationships, the rapid expansion of microbiome research presents a significant challenge: knowledge remains scattered across scientific literature, hindering comprehensive analysis and clinical translation. To address this, we introduce MINERVA (MIcrobiome NEtwork Research and Visualization Atlas), an innovative platform that leverages a fine-tuned Large Language Model to systematically map microbe-disease associations across extensive scientific literature. MINERVA constructs a rich, ontology-driven knowledge graph that prioritizes accuracy and transparency, enabling efficient exploration and discovery of previously hidden associations relevant to clinical decision-making. The platform features specialized modules that allow researchers to analyze individual microbes and diseases, visualize complex relationships within the knowledge network, uncover hidden connections through advanced graph algorithms and machine-learning models, and perform personalized and population-level microbiome compositional analysis. These capabilities facilitate the identification of disease risks, comorbidities, and actionable insights, supporting both research and clinical decision-making. By bridging the gap between microbiome research and real-world applications, MINERVA has the potential to transform our understanding of microbe-disease interactions, accelerating

discoveries and advancing patient care. MINERVA platform is available at: https://minervabio.org/

Keywords: Microbiome, Large Language Models, Knowledge Graph, Ontology, Gut Microbes.

1. INTRODUCTION

Recent advances and discoveries in biotechnology, DNA sequencing, and new computational tools have revolutionized our ability to study microbial communities in our body, leading to key discoveries about their influence on human health. Under normal conditions, the microbiota serves as a defensive barrier against pathogenic bacterial invasion, thereby bolstering the host's immune system. Additionally, it aids the host in nutrient absorption and energy derivation from food [1]. On the other hand, an imbalance in the microbiota, termed "dysbiosis," has been linked to a range of diseases. Research indicates that microbiota plays a crucial role in the onset of cardiovascular diseases, cancer, diabetes mellitus, inflammatory bowel disease, and brain disorders among others [2]. However, this critically important field poses a challenge: knowledge remains scattered across the scientific literature, hindering a comprehensive view of the current state of research and the intricate web of microbiome influences on human health. Moreover, variations in study populations, analytical techniques, and data processing biases can lead to inconsistent results when examining microbial data [3, 4]. Consequently, researchers should not rely on a single source but instead consult a diverse array of resources when investigating microbe-disease connections, making research on the field a daunting task. This underscores the pressing need for sophisticated tools capable of curating and synthesizing the expansive scientific landscape of microbiome research.

Traditional approaches like manual curation of the related scientific literature have led to the development of valuable knowledge bases such as HMDAD [5], gutMDisorder [6], Amadis [7], GMMAD [8], Disbiome [9], and MDIDB [10], among others. These resources provide a strong foundation for researchers, but their reliance on manual curation makes it challenging to keep pace with the explosive growth of microbiome research. Moreover, while these databases excel at the retrieval of curated information, they often lack capabilities to provide advanced insights, such as analyzing interactions among diverse entities or facilitating personalized microbiome analysis tailored to individual contexts.

The advent of natural language processing (NLP) and, more recently, large language models (LLMs) has introduced a transformative paradigm for extracting and synthesizing knowledge from unstructured scientific literature. By processing vast amounts of textual data, these models can automate the identification of microbe-disease associations, offering unprecedented opportunities for scalability and accuracy. However, LLMs are not without limitations. A critical concern lies in their propensity for "hallucination" whereby models generate information not grounded in evidence. This challenge is particularly acute in scientific and clinical domains, where inaccuracies could propagate misinformation with serious implications for human health [11].

To tackle these critical challenges, we introduce MINERVA (MIcrobiome NEtwork Research & Visualization Atlas), an explainable, highly accurate, and interactive knowledge platform constructed by the automated analysis of a vast corpus of scientific publications on microbe-disease associations using LLMs and a robust NLP processing pipeline, which maps the

extracted knowledge into a knowledge graph. A cornerstone of MINERVA's design is its commitment to two guiding principles: robustness and explainability. Robustness is ensured through rigorous verification processes while ingesting information, including redundancy strategies that cross-check information both within and between source documents. On the other hand, unlike conventional LLM-based systems, MINERVA grounds all its outputs in verifiable scientific evidence, enabling users to trace each finding back to its source. This combination of robustness and transparency not only minimizes the risks of hallucinations but also empowers researchers to critically assess and trust the underlying data.

MINERVA's core innovation lies in its ability to transform the fragmented landscape of microbiome research into a structured, interactive, trustworthy, and visually intuitive knowledge graph. Microbes and diseases are represented as nodes, with their relationships—derived from rigorous NLP pipelines—forming the edges. This structured representation is enriched with metadata, evidence links, and relevance scores, creating a robust foundation for advanced analyses. Furthermore, MINERVA provides a suite of interactive tools (Figure 1) that cater to the diverse needs of microbiome researchers:

- Comprehensive field analysis: Users can explore commonly studied microbes and diseases, identify research trends, and pinpoint understudied areas ripe for exploration.
- Targeted insights: Researchers can delve into specific microbe-disease relationships, accessing detailed evidence and supporting literature.
- Network analytics: Graph-based algorithms can uncover distant associations, clusters, and structural similarities, revealing hidden patterns in the data.
- Predictive modeling: MINERVA's integration with machine learning algorithms enables tasks such as link prediction and embedding similarity to forecast potential associations.
- Efficient literature synthesis: Leveraging LLM capabilities, the platform generates concise grounded summaries of research findings, highlighting key insights and knowledge gaps.
- Targeted Microbiome Health Assessment: MINERVA enables detailed comparisons between an individual's or a population's microbiome profiles and standardized healthy reference datasets. This feature identifies specific microbial imbalances and generates actionable insights by correlating deviations in microbiome abundance to specific health risks.

We envision MINERVA as a valuable resource that will empower microbiome researchers to gain deeper insights, fuel hypothesis generation, and accelerate the development of novel microbiome-based interventions to improve human health.

[Figure 1]

2. RESULTS

2.1 General overview

MINERVA is an automatically constructed knowledge base, which offers a comprehensive and up-to-date understanding of microbiome research that surpasses the scale of manually curated databases. Our automated and highly robust NLP-based pipeline has processed over 129,719 relevant publications extracted from PubMed (abstracts) or PubMed Central (complete articles when freely available), yielding a wealth of insights that surpasses the scale of

any related resource (see Figure 2). At present, MINERVA houses 3,429 microbes and 35,883 distinct diseases, directly connecting 2,941 microbes with 3,299 diseases through 66,400 distinct relationships. The remaining microbes and diseases not directly connected through explicit microbe-disease relationships are integrated into the knowledge base through hierarchical parent-child associations. This hierarchical structure enables the inference of hidden relationships between microbes and diseases, enriching the knowledge base and facilitating comprehensive exploration of potential connections within the microbiome research landscape.

MINERVA is built upon two fundamental principles: explainability and robustness. Unlike many existing resources that offer limited transparency into the evidence supporting microbe-disease relationships, MINERVA prioritizes explainability through a rigorous extraction process. This process begins with specialized Named Entity Recognition (NER) models that identify relevant microbes and diseases within each sentence of a publication. Only sentences containing both types of entities are then processed by our fine-tuned LLM, which extracts the relationship between them. Using this approach we are not only able to identify the source publication but also to pinpoint the specific sentences providing the evidence for the extracted relationship. By offering this level of granularity, MINERVA enhances user confidence in the knowledge base and enables deeper exploration of the underlying evidence.

To ensure robustness, we prioritize accuracy over exhaustive coverage by only including in our knowledge base relationships for which our fine-tuned LLM exhibits complete confidence. Furthermore, we employ a redundancy strategy that consolidates evidence from multiple mentions within and across publications, assigning a strength score to each microbe-disease relationship. This multi-layered approach not only reinforces the reliability and trustworthiness of the identified associations but also consolidates current research in case there are contradicting findings between studies. For a comprehensive explanation of these methodologies, including technical details, please refer to the Methods section.

[Figure 2]

2.2 Modules

MINERVA platform consists of several modules that can be broadly divided into three groups; (i) Exploratory (ii) Knowledge discovery and; (iii) Risk assessment modules.

2.2.1 Exploratory modules

Among the exploratory modules, the *General Statistics* module (Figure 1A) provides researchers with a high-level overview of the field, highlighting the most studied microbes and diseases, publishing trends, and key microbe-disease associations. In contrast, the *Individual Analysis* modules (Figure 1B) offer in-depth information about user-selected microbes or diseases, including basic information, research trends for that particular entity, its connections within the knowledge graph, and detailed literature evidence supporting these associations. However, for well-studied relationships (e.g., Bacillota with obesity), the volume of supporting evidence can be overwhelming. To address this, MINERVA integrates functionality for generating concise, evidence-based reports using LLMs, grounded in the platform's database. Supported models currently include GPT-4o, Gemini Pro, and AWS-hosted LLMs, accessible via API keys.

To further enhance exploration and user interaction, the *Chat Interface* module (Figure 1F) extends the capabilities of MINERVA by leveraging a chat-based LLM grounded on MINERVA's knowledge. This module allows users to ask any question about the knowledge base, whether it involves exploring specific microbe-disease relationships, retrieving information about individual entities, or even obtaining definitions and broader contextual explanations, ensuring that users are not constrained to predefined queries.

2.2.2 Knowledge Discovery Modules

While MINERVA's exploratory modules empower researchers to navigate existing knowledge, our platform is uniquely designed to also foster knowledge discovery with a suite of highly specialized modules.

By harnessing Dijkstra's shortest-path algorithm, the *Graph Algorithms* module (Figure 1C) enables the discovery of indirect relationships within the knowledge graph. Users can specify a source and target microbe or disease, and the platform identifies the shortest path, along with alternative routes, connecting these entities. These (possibly) multi-step connections, which may not be immediately apparent, can reveal hidden causal pathways and shed light on complex interactions.

The *Similarity Analysis* module (Figure 1D) enables researchers to explore the structural relationships within the knowledge graph through a variety of graph embedding algorithms. These include Node2vec [12], FastRP [13], Metapath2vec [14], and Graph Neural Network embeddings [15]. Users can select their desired embedding algorithm, visualize the representation of a chosen microbe or disease in the embedding space, and identify the most similar entities within the same class. Furthermore, the module facilitates the visualization of clusters using the K-means algorithm. Identifying microbes with similar embedding profiles may suggest shared metabolic pathways or ecological niches, while finding diseases clustered together could reveal common etiological factors or potential comorbidities [16].

Finally, the *Link Prediction* module (Figure 1F) represents a powerful tool for hypothesis generation and knowledge discovery. By leveraging our custom-trained two-layer convolutional Graph Neural Network (see Supplementary material A for training details), in this module, potential relationships between microbes and diseases that have not yet been explicitly reported in the literature are displayed. Users can select a microbe or disease of interest, and the module will display the most probable positive and negative associations for that particular entity, as predicted by our model.

2.2.3 Risk Assessment Modules

The risk assessment modules (Figure 1E and 1F) represent a key practical application of MINERVA's knowledge base, providing predictive insights into health risks through microbiome profiling. Users can upload their own or population-level microbial composition profiles and harness MINERVA's analytical capabilities to explore potential health implications. The platform compares these profiles, at a specified taxonomic level, against a control group, which can be a healthy cohort from GMRepo [17] already integrated into MINERVA or a user-defined dataset. Significant deviations in microbial abundances are automatically flagged and linked to associated diseases within MINERVA's knowledge graph. This streamlined process produces an

easily interpretable personalized or population-level risk profile, empowering users to implement proactive health management strategies or advance research into microbiome-associated health risks.

2.3 Comparison with Other Resources

To assess the robustness and accuracy of MINERVA, we compared it with five well-established, open-source, manually curated resources, as detailed in Table 1. This comparison examined overlaps in covered microbes, diseases, relationships, and their associated labels. To ensure methodological rigor, we aggregated repeated relationships in the benchmark resources using a voting scheme. Beyond direct entity matches, we also considered inferred relationships involving entities one step apart in taxonomic (microbes) or nosological (diseases) hierarchies. This hierarchical approach accommodated variations in granularity across databases, enabling a comprehensive evaluation of MINERVA's coverage and precision relative to existing resources.

From the first and second columns of Table 1, it is evident that MINERVA encompasses most of the microbes and diseases covered in the other databases. However, the percentage of overlapping relationships, as shown in the third column, tends to be lower. This is mostly due to MINERVA's emphasis on accuracy over exhaustive coverage, as our platform only incorporates relationships where our relation extraction LLM exhibits high confidence.

Notably, among the overlapped relationships, a substantial number of discrepancies were found between the labels assigned by MINERVA and those in other databases, as shown in the fourth column of Table 1. Given the scale of these discrepancies, involving thousands of relationships, we employed OpenAI's GPT-4o [18] model and Google's Gemini 1.5 Pro [19] as independent reviewers to assess the conflicting labels. Since the benchmarked resources lack specific evidence for their labels (except for MDIDB [10], which provides one sentence for each relation), we provided some of MINERVA's compiled evidence as input for the reviewers' assessment. Remarkably, as demonstrated in the last two columns of Table 1, both closed-source LLMs overwhelmingly agreed with MINERVA's labels in cases of conflict, underscoring the accuracy and robustness of our resource even in comparison to manually curated databases.

[Table 1]

This outcome suggests that MINERVA not only rivals but potentially surpasses the accuracy of manually curated databases in capturing the nuanced relationships between the microbiome and diseases within the vast body of scientific literature. This has significant implications for the field, as it highlights the potential of AI-driven curation to not only accelerate and enhance knowledge exploration in the microbiome domain but also to do so more efficiently and cost-effectively. While manual curation is labor-intensive, time-consuming, and prone to human error, AI-powered tools like MINERVA can rapidly synthesize information from massive datasets, enabling researchers to focus on interpretation and hypothesis generation. Moreover, MINERVA's inherent explainability, where supporting evidence is readily available for each prediction, empowers researchers to critically assess and build upon the existing knowledge base with increased confidence. Figure 3 shows some illustrative examples of conflicting relations between MINERVA and the other resources. For additional examples, please refer to Supplementary Material B.

[Figure 3]

2.4 Case Study: Population with Alzheimer's Disease

As an example of MINERVA's ability to effortlessly translate microbiome research into actionble clinical insights, we analyzed a cohort with Alzheimer's disease presented in [20] available in GMRepo [17] using the *Population Risk Assessment Module*. For this demonstration, the control group from the same study was used as a reference for comparative purposes. Is worth mentioning that the related publication of this study is not among the analyzed publications in MINERVA.

As depicted in Figure 4, after uploading microbiome composition data for both the target and control groups, the module delivers three types of results: (i) Statistical Analysis: Metrics such as α-diversity and β-diversity are computed to assess bacterial diversity within and between groups. Additionally, Partial Least Squares-Discriminant Analysis (PLS-DA) [21] is employed to identify microbes that most significantly discriminate the groups. (ii) Disease-Specific Microbial Imbalances: When a target disease is specified, MINERVA can highlight relevant microbial imbalances associated with an increased or decreased risk of that particular condition. In this case, as shown in Figure 4B, Bacteroides, Roseburia, and Ruminococcus are identified as the most important microbial imbalances associated with Alzheimer's disease in the analyzed population. (iii) Risks of diseases: Using only the microbial imbalances of each individual, MINERVA is able to compute disease risk scores based on its extensive knowledge graph. These individual risk scores are then aggregated to identify diseases with the highest overall risk. Notably, as shown in Figure 4C, even without specifying Alzheimer's disease as the target, it emerged as one of the highest-ranked diseases in the population analysis. This demonstrates MINERVA's diagnostic capabilities, leveraging microbial data to uncover latent disease associations and providing a powerful tool for population-level risk assessment.

For a detailed example showcasing the usage and outcomes of the *Individual Risk Assessment* module, and a more detailed example showcasing the results of the *Population Risk Assessment Module* the reader is referred to Supplementary Material C and D, respectively.

[Figure 4]

3. MATERIALS AND METHODS

The construction of MINERVA followed a systematic four-step pipeline (Figure 5): (A) Data collection from PubMed and PubMed Central, (B) Data processing using specialized named entity recognition (NER) models and LLMs to extract meaningful microbe-disease relationships, (C) Construction of a robust knowledge graph, and (D) Development of an intuitive web-based platform to enable seamless exploration, analysis, and hypothesis generation.

[Figure 5]

3.1 Data Collection

To populate MINERVA, we collected 129,719 relevant PubMed abstracts published between 2014 and 2023, obtained by the search query *((microbiome) OR (dysbiosis) OR (microbiome*

alterations)) AND (``2013"[Date - Create]: ``2023"[Date - Create]) (see Figure 4A). For those abstracts where the full paper was freely available on PubMed Central (PMC), we obtained the full paper as well, resulting in 78,382 full papers.

As depicted in Figure 2, the number of published microbiome studies has grown substantially in recent years. This rapid expansion of research in this field underscores the need for automated solutions like MINERVA, empowering researchers to efficiently navigate and leverage the latest discoveries in this dynamic area.

3.2 Data Processing

For each relevant publication (abstract or full paper), we split it into sentences and analyzed each sentence independently using two named entity recognition (NER) models, one trained to identify microbes and the other trained to identify diseases. If a sentence contains at least one microbe and at least one disease, we used our fine-tuned LLM to infer the relationship between the microbe(s) and disease(s) present in the sentence (Figure 5B). Sentences lacking either entity were excluded from further analysis.

3.2.1 Name Entity Recognition

Unlike previous studies that have used dictionary-matching techniques for disease and microbe entity recognition from scientific publications [10, 22], in this work we adopted transformers-based NER models based on the BERT architecture. due to its superior performance and generalization [23].

Specifically, we used the SciBERT model [24] for disease NER, and a custom microbial NER solution by finetuning the DistilBERT model [25] on the BNER2.0 dataset [26]. Using this model we achieved, an F1 score of 0.914, a precision of 0.951, and a recall of 0.895 over the test set. Following entity identification, we employed ScispaCy's Entity Linker [27] for normalization, assigning a UMLS Concept Unique Identifier (CUI) to each recognized entity to ensure standardized representation and facilitate downstream analysis.

3.2.2 Relation extraction

Relation extraction was performed using a fine-tuned GPT-based LLM, specifically using its ability to understand nuanced scientific language and complex relationships within the text [28]. We fine-tuned and evaluated the model using the dataset provided by [29], a collection of 1,100 manually labeled sentences categorizing microbe-disease relationships as Positive (microbe promotes the disease), Negative (microbe inhibits the disease), Related (unclear direction of the relation), or NA (unrelated). To improve clarity and focus on the directionality of the relationships, which is an important feature of MINERVA, we merged the Related and NA labels into a single Unrelated category. This resulted in a refined dataset with 571 Positive, 305 Negative, and 224 Unrelated sentences.

Using this dataset, we compared the performance of several open-source fine-tuned and non-fine-tuned models under zero-shot and few-shot conditions, employing a five-fold validation approach. Since [29] found BERT-based models to perform competitively with or even surpass

GPT-based models on this task, we included them in our evaluation. Closed-source LLMs were excluded due to the prohibitive cost of applying them to large-scale scientific literature analysis.

In the case of GPT-based models, we applied self-consistency [30], which usually consists of sampling multiple answers with varying temperatures and prioritizing answers achieving a majority vote. While the majority approach is useful in general multiple-choice tasks, for our high-confidence knowledge graph construction, we just included relations with complete consistency. The remaining relations were discarded, prioritizing accuracy over wide coverage.

Table 2 reveals that few-shot learning substantially improves non-finetuned models. However, fine-tuning, even on smaller models, consistently outperforms non-finetuned approaches. This is likely due to the highly technical scientific terminology prevalent in microbiome-related publications, which fine-tuning helps the models to better understand and process. Among the finetuned models, the Biomistral 7b model [31] achieved the best trade-off between accuracy and coverage for full-confidence (F.C.) predictions. We then proceeded to boost its performance even further by using a data-augmentation approach. This approach leveraged the Mixtral 8x7b LLM [32], prompting it to generate multiple rephrased versions of sentences from our database that described the relationship between a given microbe and disease. Then for these rephrased sentences, we replaced the original microbe and disease by randomly chosen entities from the original database. We used these additional sentences to augment the database and train the model with them. Training on this augmented dataset yielded Biomistral-AUG 7b, which demonstrates the best overall performance and the best trade-off between accuracy and coverage for full-confidence predictions. The specific prompt used for the zero-shot, few-shot, and fine-tuning approaches is presented in the supplementary material E.

[Table 2]

3.3 Robust knowledge graph construction

In addition to only considering the high confidence relations predicted by our LLM (Figure 6A), to further ensure the precision of our knowledge graph, we employed a two-tiered approach. First, we addressed potential discrepancies within a single paper, that is, for a microbe-disease pair discussed in multiple sentences within the same paper, if those sentences presented conflicting relationship labels (Positive, Negative, or Unrelated), a majority voting approach determined the final label (Figure 6B). Prioritizing clarity and directionality, we retained only Positive and Negative relationships, discarding Unrelated ones. This ensures our knowledge graph captures only strongly verified associations, which are then incorporated as triplets (m,r,d). Here, m and d represent microbe and disease nodes respectively, and r is the relationship between them, represented as an edge in the knowledge graph.

Secondly, to resolve potential conflicts in microbe-disease relationships across multiple papers, we implemented a weighting approach, that consists of the following: Given N relationships extracted from different papers between microbe m and disease d, we calculate the strength of the association between them as follows:

$$w_{md} = \sum_{k=1}^{N} r_{k_{mdp}} \quad (1)$$

With $r_{k_{mdp}}$ as the strength of the relationship k between m and d coming from paper p, which is given by:

$$r_{k_{mdp}} = sign \cdot log_{10}(F_p) \quad (2)$$

Where F_p is the impact factor of the journal in which paper p was published and $sign$ is +1 if the relationship is labeled as positive and -1 if the relationship is labeled as negative. We slightly favor papers published in more prestigious journals, as these publications generally undergo a more rigorous peer-review process, potentially increasing the reliability of their findings (Figure 6C).

Finally, we enriched the knowledge graph by incorporating hierarchical relationships. Disease entities were expanded using the SNOMED CT ontology, while taxonomical structures were applied to microbes, adding parent and child nodes. This enrichment broadens the graph's scope and enhances its utility for analyzing multi-hop microbe-disease connections across hierarchical trees.

[Figure 6]

3.4 MINERVA's Interface

Finally, for the practical implementation of the knowledge graph we used Neo4j [33] and to create MINERVA's user-friendly interface, we employed the Streamlit framework [34], which allows for the rapid development of interactive, data-driven web applications. MINERVA's platform is freely available at https://minervabio.org/, and a video showcasing its use at https://drive.google.com/drive/folders/1ODtXOe7op1A2dIil3xVHwfkItyH7Mzx8?usp=sharing.

4. DISCUSSION

MINERVA represents a valuable tool for the field of microbiome research, providing a comprehensive, explainable, and up-to-date platform that navigates the complex and rapidly expanding landscape of known and sometimes conflicting findings on microbiome-disease associations. By integrating advanced natural language processing techniques, rigorous consistency checks, graph algorithms, and machine learning models, MINERVA provides an intuitive and interactive interface that enables researchers to efficiently conduct systematic reviews, explore vast datasets, and uncover hidden patterns. This capability not only facilitates the generation of novel hypotheses but also accelerates the pace of discovery in this rapidly evolving field. By synthesizing diverse and sometimes conflicting findings, MINERVA offers a valuable resource for deriving actionable insights, with the potential to inform and enhance both research initiatives and clinical practices.

For instance, as highlighted in the case study of the population with Alzheimer's disease, MINERVA could enable researchers to rapidly discern hidden associations, such as the roles specific microbes may play in the progression or mitigation of cognitive impairments across the spectrum from mild cognitive impairment to dementia. In clinical practice, these findings could guide physicians in modulating an individual's microbiome to counteract patterns associated with dementia and foster those linked to healthier conditions through targeted medication, dietary interventions, and supplementation. This underscores the vital role of microbiome research in advancing personalized medicine while showcasing MINERVA's capability to distill

complex datasets into practical, actionable insights. By bridging the gap between raw scientific data and its application in clinical settings, MINERVA could serve as a catalyst for transforming microbiome research into meaningful advancements in patient care.

Despite its potential, MINERVA has some limitations that warrant consideration. The database mainly relies on abstracts and open-access papers from PubMed, potentially biasing its scope. While our AI knowledge extraction models and redundancy strategy are highly accurate, the current weighting system, prioritizing results from high-impact journals, may be unstable due to fluctuating impact factors. Additionally, the pipeline focuses on single-sentence relationships, possibly missing complex associations spanning multiple sentences or paragraphs. Finally, MINERVA's insights are based on existing research and lack validation through direct microbiome analysis, necessitating empirical studies to confirm its findings.

To address these limitations and advance MINERVA, we will pursue several strategic directions. First, we aim to expand the knowledge base by incorporating additional repositories beyond PubMed and data sources such as clinical trial outcomes, longitudinal metagenomics, and metabolomics. This will reduce bias and foster a holistic understanding of microbiome compositions and health implications. Second, we plan to enhance the natural language processing pipeline by implementing cross-sentence relationship extraction and coreference resolution to capture complex microbiome-disease associations spanning multiple sentences. Additionally, we will develop a refined weighting strategy that accounts for factors like study size, methodological rigor, and replication status. These improvements will transform MINERVA into a more powerful tool, generating nuanced insights while maintaining accuracy and transparency.

Another promising future direction for our work is the integration of additional entities into our knowledge graph, following a similar methodology to that described in this work. These entities could include the effects of diet and nutrition, environmental exposures, lifestyle factors such as physical activity and stress, and pharmaceutical interventions on the microbiome. By broadening the scope of MINERVA to encompass these dimensions, the platform could provide a more holistic understanding of the interplay between the microbiome and human health.

In conclusion, MINERVA represents a significant step forward in microbiome research and clinical application. This comprehensive, explainable, and interactive platform leverages an extensive database and cutting-edge computational tools to systematically map and analyze microbiome-disease associations through a robust knowledge graph. By providing researchers and clinicians with streamlined access to microbiome patterns linked to specific diseases, MINERVA facilitates the identification of actionable insights, enabling interventions to prevent comorbidities or mitigate disease progression. Furthermore, the platform’s ability to continuously evolve by incorporating new findings ensures its adaptability and relevance as the field of microbiome research expands. While challenges remain, MINERVA provides a valuable tool for advancing our understanding of the microbiome’s role in human health. With further refinement and broader integration, MINERVA may serve as an indispensable resource for both clinical applications and microbiome research, contributing to progress in precision medicine and public health.

Key Points

1. **Addressing Scattered Knowledge:** Microbiome research is rapidly expanding, but scattered knowledge across scientific literature hinders comprehensive understanding and clinical application.
2. **AI-Powered Knowledge Graph:** MINERVA uses advanced AI techniques to create a scalable knowledge graph that maps microbe-disease associations from over 129,000 publications.
3. **Ensuring Accuracy and Transparency:** The platform prioritizes accuracy through rigorous verification processes and allows users to trace information back to original sources.
4. **Tool for Research and Clinical Insights:** MINERVA offers tools for exploring trends, discovering hidden relationships, and predicting new associations, with potential to accelerate research and inform clinical decision-making.

Funding:

This work was supported by HEM Pharma; and AWS Health Equity Initiative, which provided credits to use their computational infrastructure to finetune the LLMs used in this study.

Acknowledgments:

The authors express their sincere gratitude to Michelle Chua for her initial contributions to this project and to Kyungsu Kim for his suggestions in the initial stages of this project. We also acknowledge the NIH for providing access to PubMed and PubMed Central.

Declaration of interests:

Authors declare that they have no competing interests

Author contributions:

SL carried out the experiments and wrote most of the manuscript. YTK designed and created visual representations of the study materials and made contributions by carefully reviewing and revising the manuscript. AA trained the entity recognition models and revised the manuscript thoroughly. JBK provided critical feedback, rigorously reviewed the manuscript, and contributed to its revision. SD study conception and design, continued advice, and revision of the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Data availability statement (DAS):

All data, and code for reproducibility is available at https://github.com/MGH-LMIC/MINERVA , additionally MINERVA's platform is freely available at https://minervabio.org/.

References

1. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006 Dec;444(7122):1027-31.
2. Hadrich D. Microbiome Research Is Becoming the Key to Better Understanding Health and Nutrition. Front Genet. 2018 Jun;9:212.
3. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, et al. Microbiome differential abundance methods produce different results across 38 datasets. Nature Communications. 2022 Jan;13(1):342.
4. Nearing JT, Comeau AM, Langille MGI. Identifying biases and their potential solutions in human microbiome studies. Microbiome. 2021 May;9(1):113.
5. Ma W, Zhang L, Zeng P, Huang C, Li J, Geng B, et al. An analysis of human microbe-disease associations. Brief Bioinform. 2016 Feb;18(1):85-97.
6. Qi C, Cai Y, Qian K, Li X, Ren J, Wang P, et al. gutMDisorder v2.0: a comprehensive database for dysbiosis of gut microbiota in phenotypes and interventions. Nucleic Acids Research. 2022 10;51(D1):D717-22.
7. Li L, Jing Q, Yan S, Liu X, Sun Y, Zhu D, et al. Amadis: A Comprehensive Database for Association Between Microbiota and Disease. Front Physiol. 2021 Jul;12:697059.
8. Wang CY, Kuang X, Wang QQ, Zhang GQ, Cheng ZS, Deng ZX, et al. GMMAD: a comprehensive database of human gut microbial metabolite associations with diseases. BMC Genomics. 2023 Aug;24(1):482.
9. Janssens Y, Nielandt J, Bronselaer A, Debunne N, Verbeke F, Wynendaele E, et al. Disbiome database: linking the microbiome to disease. BMC Microbiol. 2018 Jun;18(1):50.
10. Wu C, Xiao X, Yang C, Chen J, Yi J, Qiu Y. Mining microbe–disease interactions from literature via a transfer learning model. BMC Bioinformatics. 2021 Sep;22(1):432.
11. Aurangzeb Ahmad M, Yaramis I, Dutta Roy T. Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI. arXiv e-prints. 2023 Sep:arXiv:2311.01463.
12. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 855–864.
13. Chen H, Sultan SF, Tian Y, Chen M, Skiena S. Fast and Accurate Network Embeddings via Very Sparse Random Projection. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 399–408.
14. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable Representation Learning for Heterogeneous Networks.In: Proceedings of the 23rd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining. KDD '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 135–144.

15. Wu W, Li B, Luo C, Nejdl W. Hashing-Accelerated Graph Neural Networks for Link Prediction. In: Proceedings of the Web Conference 2021. WWW '21. New York, NY, USA: Association for Computing Machinery; 2021. p.2910–2920.
16. Fu C, Zhong R, Jiang X, He T, Jiang X. An Integrated Knowledge Graph for Microbe-Disease Associations. In: Huang Z, Siuly S, Wang H, Zhou R, Zhang Y, editors. Health Information Science. Cham: Springer International Publishing; 2020. p. 79-90.
17. Dai D, Zhu J, Sun C, Li M, Liu J, Wu S, et al. GMrepo v2: a curated human gut microbiome database with special focus on disease markers and cross-dataset comparison. Nucleic Acids Res. 2022 Jan;50(D1):D777-84.
18. OpenAI: GPT-4 Technical Report (2024). https://doi.org/10.48550/arXiv.2303.08774.
19. Gemini: A Family of Highly Capable Multimodal Models (2024). https://doi.org/10.48550/arXiv.2312.11805.
20. Liu P, Wu L, Peng G, Han Y, Tang R, Ge J, et al. Altered microbiomes distinguish Alzheimer's disease from amnestic mild cognitive impairment and health in a Chinese cohort. Brain, Behavior, and Immunity. 2019;80:633-43.
21. Pérez-Enciso M, Tenenhaus M. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Human Genetics. 2003 May;112(5):581-92.
22. Park Y, Lee J, Moon H, Choi YS, Rho M. Discovering microbe-disease associations from the literature using a hierarchical long short-term memory network and an ensemble parser model. Scientific Reports. 2021 Feb;11(1):4490.
23. Nastou K, Koutrouli M, Pyysalo S, Jensen LJ. Improving dictionary-based named entity recognition with deep learning. bioRxiv. 2023.
24. Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 3615-20.
25. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv e-prints. 2019 Oct:arXiv:1910.01108.
26. Li X, Fu C, Zhong R, Zhong D, He T, Jiang X. Bacterial Named Entity Recognition Based on Language Model. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2019. p. 2715-21.
27. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Florence, Italy: Association for Computational Linguistics; 2019. p. 319-27.
28. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. Nature Reviews Physics. 2023 May;5(5):277-80.

29. Karkera N, Acharya S, Palaniappan SK. Leveraging pre-trained language models for mining microbiome-disease relationships. BMC Bioinformatics. 2023 Jul;24(1).

30. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., et al.: Towards Expert-Level Medical Question Answering with Large Language Models. arXiv e-prints, 2305–09617 (2023).

31. Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. arXiv e-prints. 2024 Feb:arXiv:2402.10373.

32. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of Experts. arXiv e-prints. 2024 Jan:arXiv:2401.04088.

33. Neo4j, Inc . Neo4j Graph Database; 2023. Accessed: March 28, 2024. Available from: https://neo4j.com/.

34. Streamlit; 2024. Accessed: March 28, 2024. Available from: https://streamlit.io/.

35. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthcare. 2021 oct;3(1).

36. Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining Language Models with Document Links. In: Association for Computational Linguistics (ACL); 2022.

37. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of Experts. arXiv e-prints. 2024 Jan:arXiv:2401.04088.

38. Touvron, H., Martin, L., Stone, K., Albert, P., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv e-prints*, arXiv:2307.09288.

39. Mitra A, Del Corro L, Mahajan S, Codas A, Simoes C, Agarwal S, et al. Orca 2: Teaching Small Language Models How to Reason. arXiv e-prints. 2023 Nov:arXiv:2311.11045.

40. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Singh Chaplot D, de las Casas D, et al. Mistral 7B. arXiv e-prints. 2023 Oct:arXiv:2310.06825.

41. Tunstall L, Beeching E, Lambert N, Rajani N, Rasul K, Belkada Y, et al. Zephyr: Direct Distillation of LM Alignment. arXiv e-prints. 2023 Oct:arXiv:2310.16944.

Tables

Database	Overlapping Microbes	Overlapping Diseases	Overlapping Relationships	Differing labels with MINERVA	GPT4o coincidence with MINERVA/	Gemini's coincidence with MINERVA/

					other DB [%]	other DB [%]
AMADIS (*7*)	683 (87%)	206 (82%)	2375 (77%)	1069 (45%)	88.8% / 3.7%	89.2% / 3.7%
GMMAD (*8*)	359 (69%)	108 (97%)	1516 (65%)	636 (42%)	85.3% / 3.7%	85.3% / 3.7%
HMDAD (*5*)	210 (72%)	35 (91%)	274 (61%)	60 (22%)	86.2% / 6.9%	89.7% / 5.2%
DISBIOME (*9*)	1141 (72%)	322 (90%)	5521 (63%)	1932 (35%)	84.5% / 3.9%	84.3% / 4.3%
MDIDB (*10*)	284 (90%)	302 (73%)	897 (92%)	269 (30%)	92.2% / 3.6%	93.4% / 4.8%

Table 1. Comparison of MINERVA with other resources. Resources are compared in terms of coverage and accuracy. Since there are hundreds of relationships accuracy is assessed using closed-source state-of-the-art LLMs.

Model	Accuracy	F1-Score	Accuracy (F.C.)	F1-Score (F.C.)	Coverage (F.C.)
PubMed-BERT (Finetuned) [35]	0.790	0.777	-	-	-
BioLink-BERT (Finetuned) [36][1]	0.797	0.787	-	-	-

[1] Best performing model in (*29*)

Mixtral 8x7b (Zer-Shot) [37]	0.610	0.636	0.695	0.716	59%
Mixtral 8x7b (Few-Shot)	0.801	0.782	0.832	0.812	91%
Llama 2 70b (Zero-Shot) [38]	0.217	0.110	0.213	0.08	66%
Llama 2 70b (Few-Shot)	0.765	0.741	0.805	0.778	86%
Orca 13b (Zero-Shot) [39]	0.658	0.603	0.715	0.655	73%
Orca 13b (Few-Shot)	0.673	0.634	0.751	0.713	68%
Mistral 7b (Finetuned) [40]	0.818	0.806	0.907	0.893	66%
Zephyr 7b (Finetuned) [41]	0.828	0.814	0.920	0.902	65%
BioMistral 7b (Finetuned) [31]	0.824	0.808	0.881	0.859	78%
BioMistral-AUG-7b (Finetuned)	0.847	0.841	0.890	0.884	81%

Table 2. Comparison of different open-source LLMs for relation extraction. The comparison is done using a Five-fold validation approach on (*29*) dataset. The objective is to find the best possible trade-off between accuracy and coverage of full confidence (F.C.) prediction.

Figures

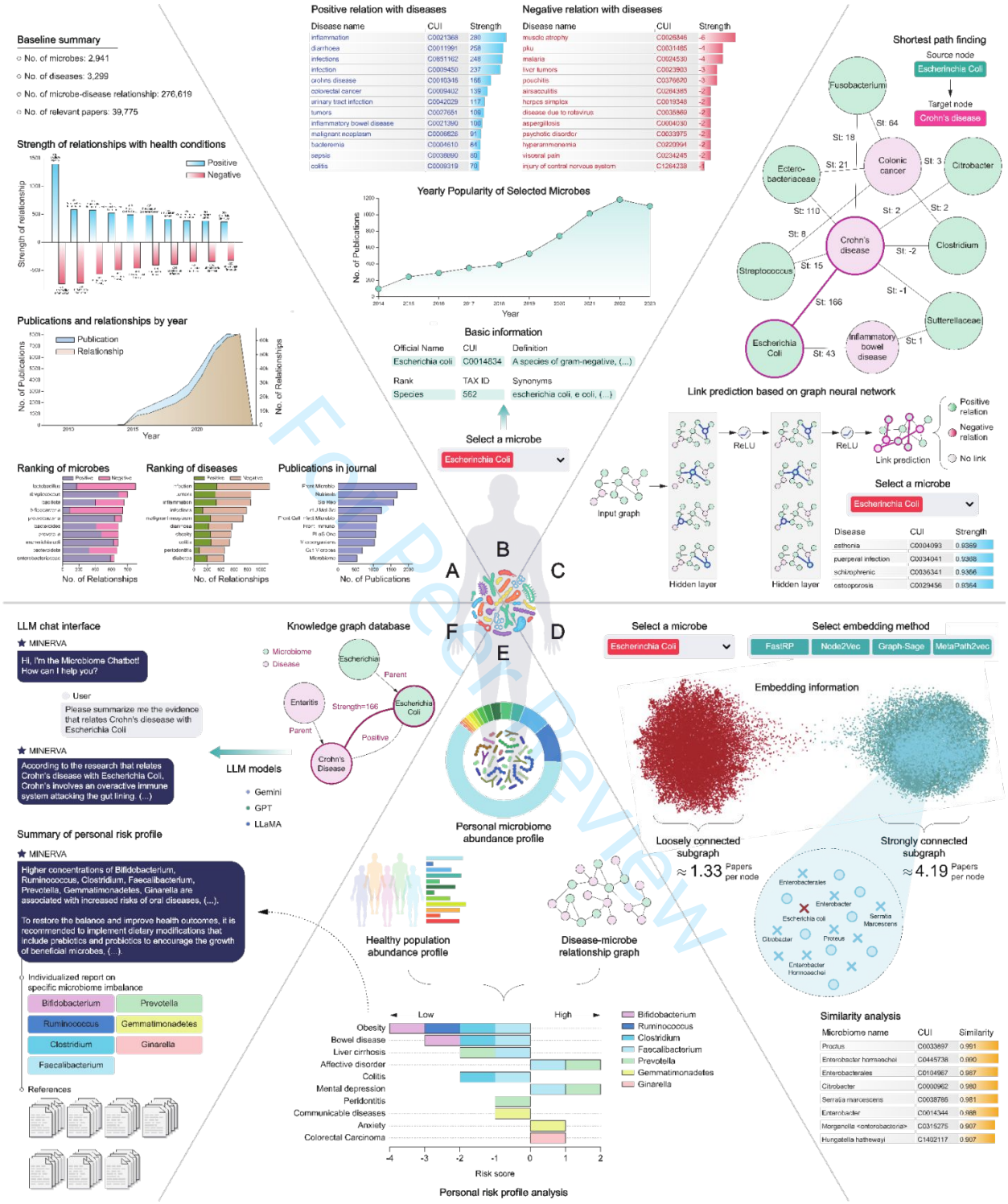

Fig. 1. Overview of MINERVA's tools for exploring and analyzing microbe-disease relationships. (A) The General Statistics module provides a summary of the database, highlighting frequently studied microbes and diseases, common relationships, and key journals. (B) The Entity Analysis modules offer a comprehensive view of individual entities (microbes or diseases). (C) The Link Prediction and Path Finding modules identify novel microbe-disease associations using a custom-trained Graph Neural

Network and reveal potential multi-hop connections between entities to uncover hidden relationships. (D) The Similarity Analysis module visualizes entity embeddings, enables clustering through K-means, and identifies similar entities based on embedding similarity scores. (E) The Personal and Population Microbiome Analysis modules compare microbiome abundance profiles to that of a healthy population. Microbes with abundances outside the healthy range are flagged and disease risk scores are computed. (F) The Chat Interface module enables researchers to engage interactively with the platform by the use of a conversational large language model integrated with the knowledge graph.

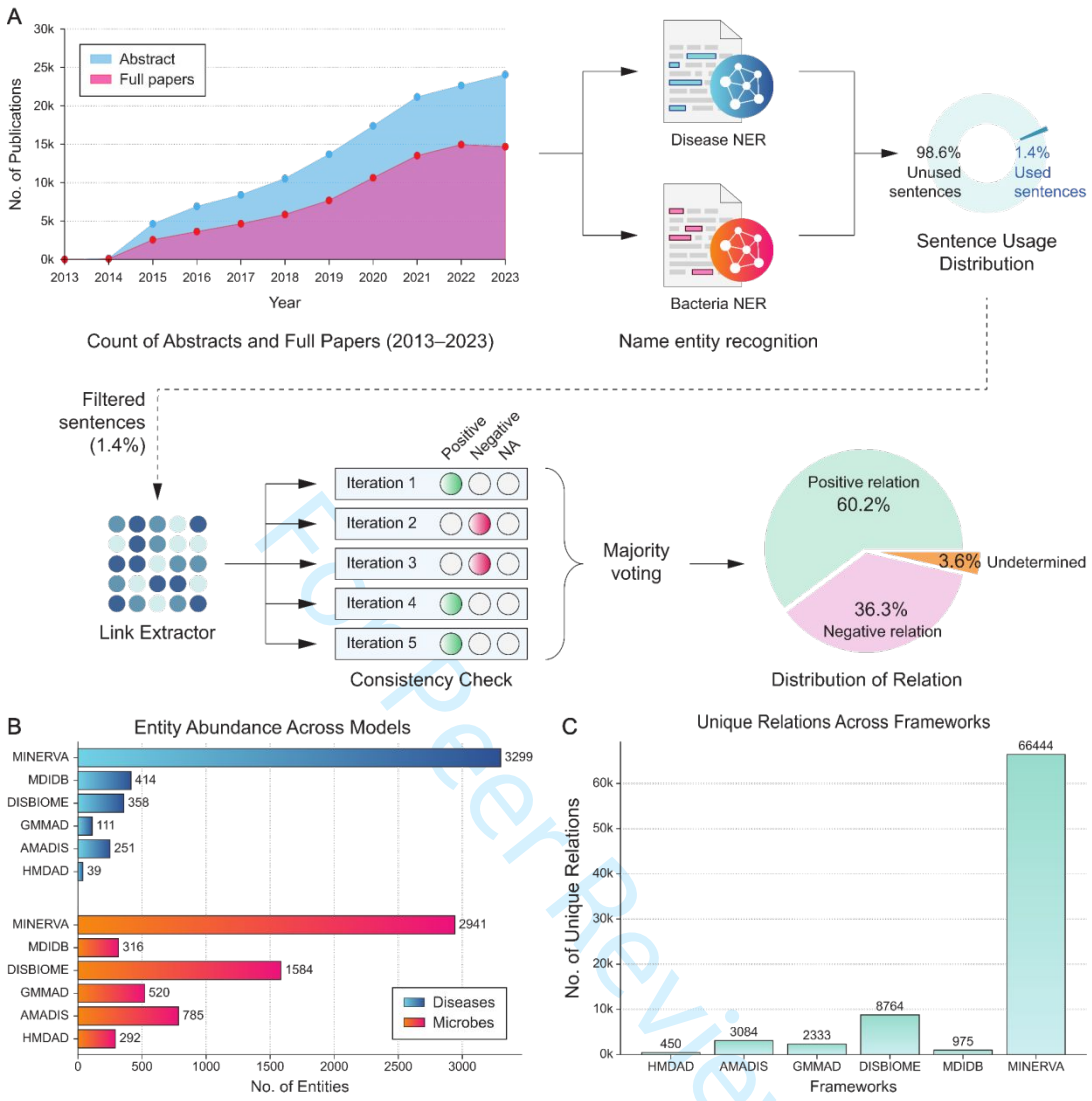

Fig. 2. Statistics and general overview of MINERVA's data processing pipeline. (A) MINERVA has processed 129,719 microbiome-related publications from PubMed and PubMed Central, spanning the years 2014 to 2023. These publications were analyzed using two specialized BERT-based NER models for microbe and disease entity recognition. Only 1.4% of sentences containing both entities were selected for further processing by a fine-tuned GPT-based LLM to extract relationships. The resulting relationships were validated through consistency checks and majority voting, leading to a label distribution of 60.2% positive, 36.3% negative, and 3.6% undetermined. (B) MINERVA includes 3,299 diseases and 2,941 microbrews, far exceeding the scope of any related resource. (C) Unique relationships extracted across frameworks are compared, with MINERVA achieving a total of 66,444 distinct microbe-disease relations, far surpassing existing resources.

Resource	Microbe	Disease	Resource Label	Resource Evidence	MINERVA Label	MINERVA Evidence	
						Pubmed ID	Sentence
AMADIS	Enterobacteriaceae	I. Bowel Disease	NEGATIVE	Pubmed ID 28683448	POSITIVE	26374288	Investigating individuals with **inflammatory bowel disease**, knights et al. have shown that nod2 risk allele count is correlated. with an increase in the relative abundance of **enterobacteriaceae**
						35316142	inflammatory conditions of the intestinal tract, such as **inflammatory bowel disease**, have been associated with an increased abundance of facultative anaerobes belonging to the class of gammaproteobacteria, mostly **enterobacteriaceae**
GMMAD	Bifidobacterium	Diarrhea	POSITIVE	Pubmed ID 28439072	NEGATIVE	34206053	The reduction in lactobacilli and **bifidobacteria** have both been linked to **diarrhea**, food allergy, and chronic inflammatory bowel diseases
						30836671	Introduction of probiotic strains of **bifidobacteria** to human gut has been reported to improve clinical status in diseases like antibiotic associated **diarrhea** [7, 8], necrotizing enterocolitis [9], chronic pouchitis [10].
DISBIOME	Bacteroides	Obesity	POSITIVE	DOI 10.1038/ s41598-019-49462-w	NEGATIVE	29324644	A decrease in **bacteroides** is suggested to be associated with metabolic diseases, such as **obesity** and diabetes [30, 31].
						34201465	a reduction in **bacteroides** is associated with the pathological onset of **obesity**, and the increase in lactobacillus is a biomarker of irreversible pathophysiological phenomena [116]
HMDAD	Diabetes Mellitus	Porphyromonas	NEGATIVE	Pubmed ID 20613793	POSITIVE	37511934	The pathogenic bacteria responsible for periodontitis, such as **porphyromonas gingivalis** and prevotella intermedia, were found to be significantly higher in patients suffering from type 1 **diabetes mellitus** when compared with periodontally healthy controls [5].
						37488251	In particular, metagenomic analysis highlighted the persistence of the gram negative bacteria treponema denticola, tannerella forsythia and **porphyromonas gingivalis**, which are part of the so called "red complex" bacteria, known to be strongly associated with periodontal disease 32, 33, but also to a higher risk of developing esophageal cancer 59, 60, and **diabetes mellitus** 61, 62
MDIDB	Fusobacterium	Cancer	NA	Pubmed ID 28063002*	POSITIVE	36046741	The virulence factors produced by **fusobacterium**, such as adhesins, Ips, and radd, have been associated with aberrant immune responses, chronic infection, modulating oral carcinogenesis, and promoting **cancer** progression (gholizadeh et al., 2017; de andrade et al., 2019).
						37566024	Moreover, **cancer** development also correlated with an augmented presence of **fusobacterium**, atopobium, gluconacetobacter, hydrogenophaga, and lactobacillus genera [80].

*The candidate reason for observed poor prognosis may due to the inverse association between Fusobacterium nucleatum and the infiltrated T-cell amount, which has a negative correlation with cancer, in the cancer sites.

Fig. 3. Conflicting labeling and comparison of available evidence between MINERVA and other resources. Unlike most related resources (except for MDIDB, which provides only one piece of evidence), MINERVA provides multiple pieces of evidence, down to the sentence level, for each microbe-disease relationship label. This enables users to easily assess the confidence and validity of the assigned label. In the Figure, only two pieces of evidence are shown for each conflicting label, however, some relations can have hundreds of supporting sentences coming from different publications.

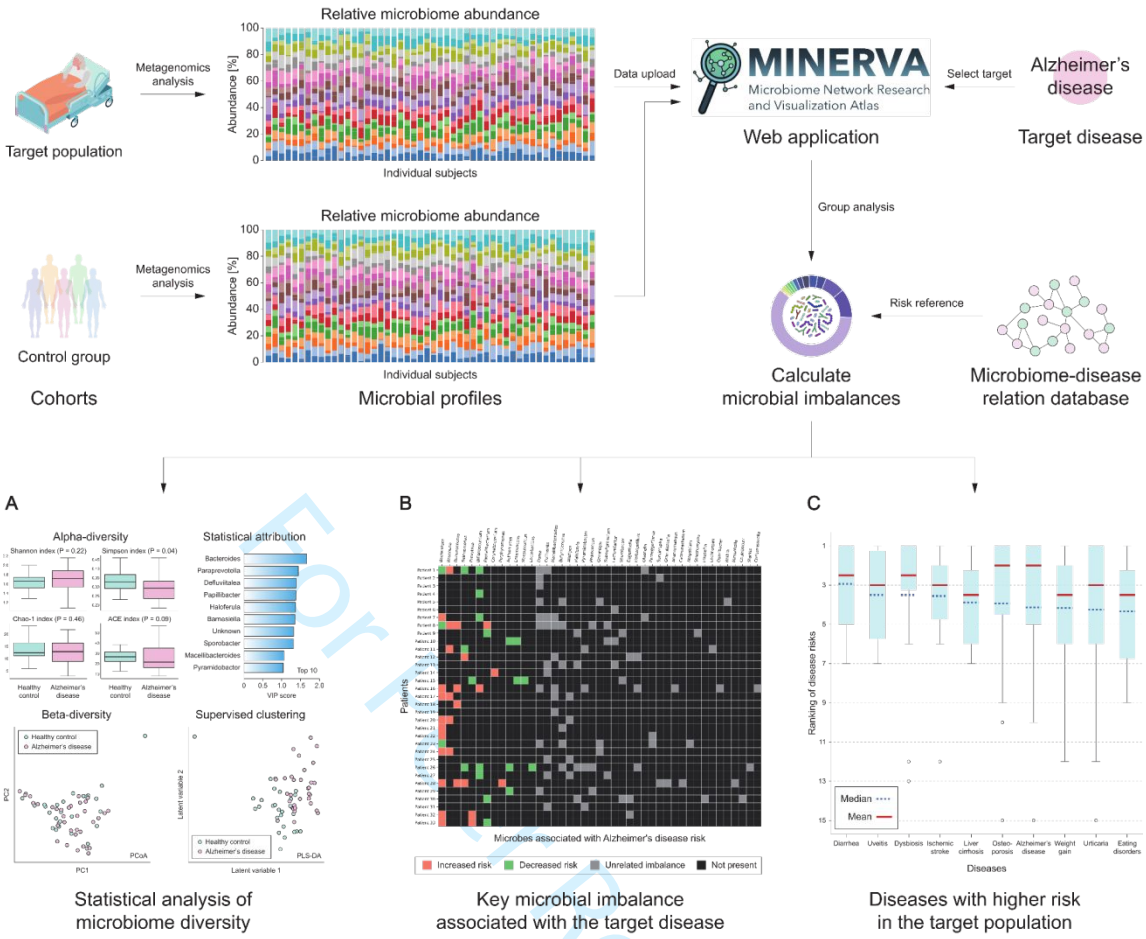

Fig. 4. Population risk assessment module. After uploading the microbial profiles of a target population and (optionally) of a control group. MINERVA calculates the microbial imbalances of the target population and delivers multiple insights, which can be broadly divided into three: (A) Statistical Analysis: Metrics such as α-diversity and β-diversity, along with methods like PLS-DA, are used to compare microbial diversity and identify key discriminative microbes between groups. (B) Disease-Specific Microbial Imbalances: If a target disease is specified, MINERVA will automatically highlight which microbial imbalances are associated to the target disease, providing actionable insights into the microbiome-disease relationship. (C) Risk of Diseases: MINERVA calculates and ranks disease risks across the population using its knowledge base, aiding in the identification of potential comorbidities and enhancing diagnostic accuracy.

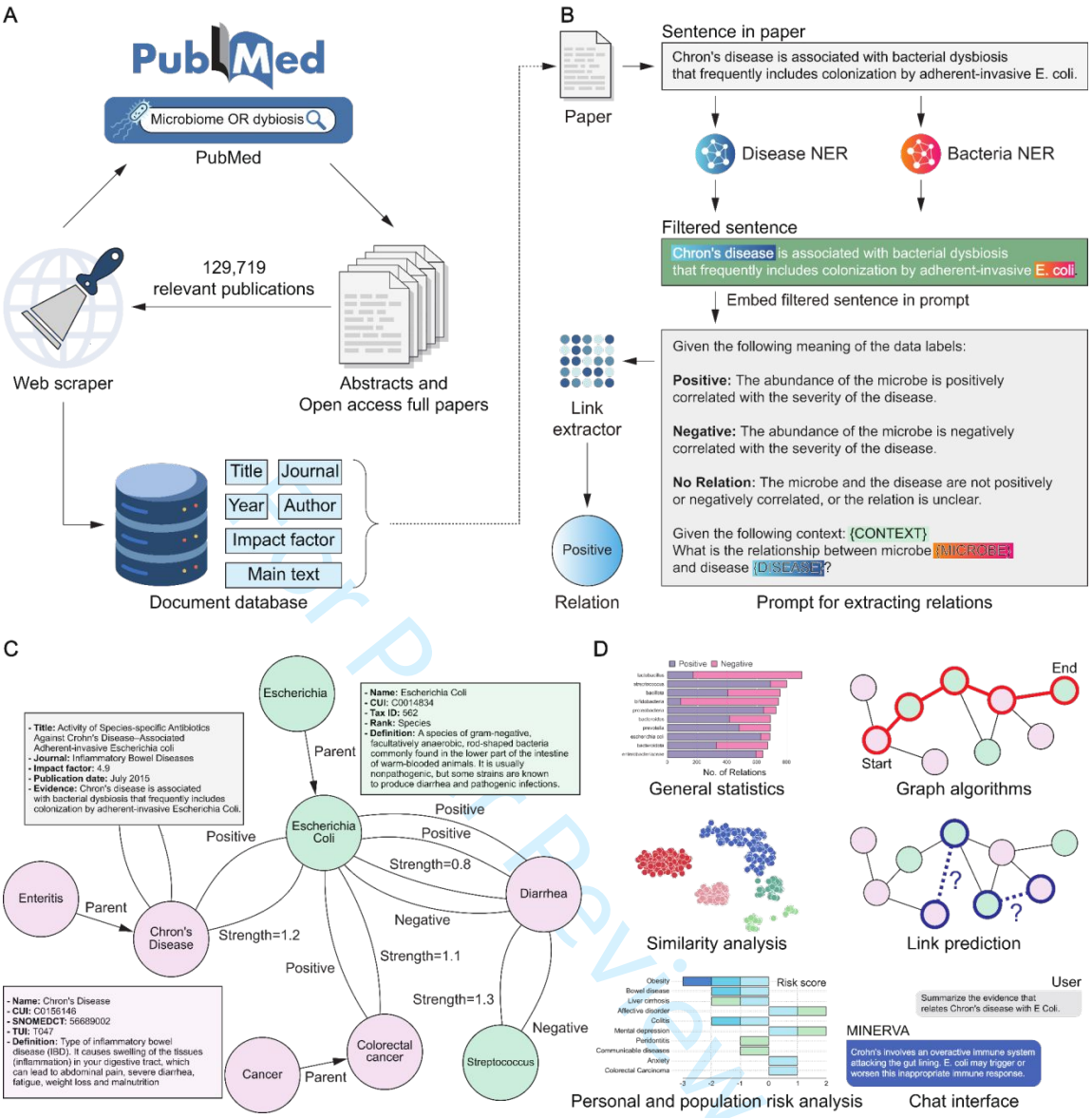

Fig. 5. Full pipeline for the construction of MINERVA. (A) Abstracts and full papers (when available) are obtained from PubMed and PubMed Central, resulting in 129,719 relevant publications. (B) Each publication is then segmented into sentences, which are processed by two specialized NER models, one for recognizing diseases and the other for recognizing microbes. Filtered sentences that include co-occurring disease and microbe entities are then embedded into a GPT-based LLM prompt to extract relationships. (C) Once microbe-relation-disease triplets are obtained for each sentence and the consistency checks are passed, the triplets are incorporated into our knowledge graph. (D) A user-friendly platform with several modules is built upon the knowledge graph. This platform empowers researchers to explore the current landscape of microbiome research, uncover novel insights, and generate hypotheses through integrated large language models, machine learning models, and graph algorithms.

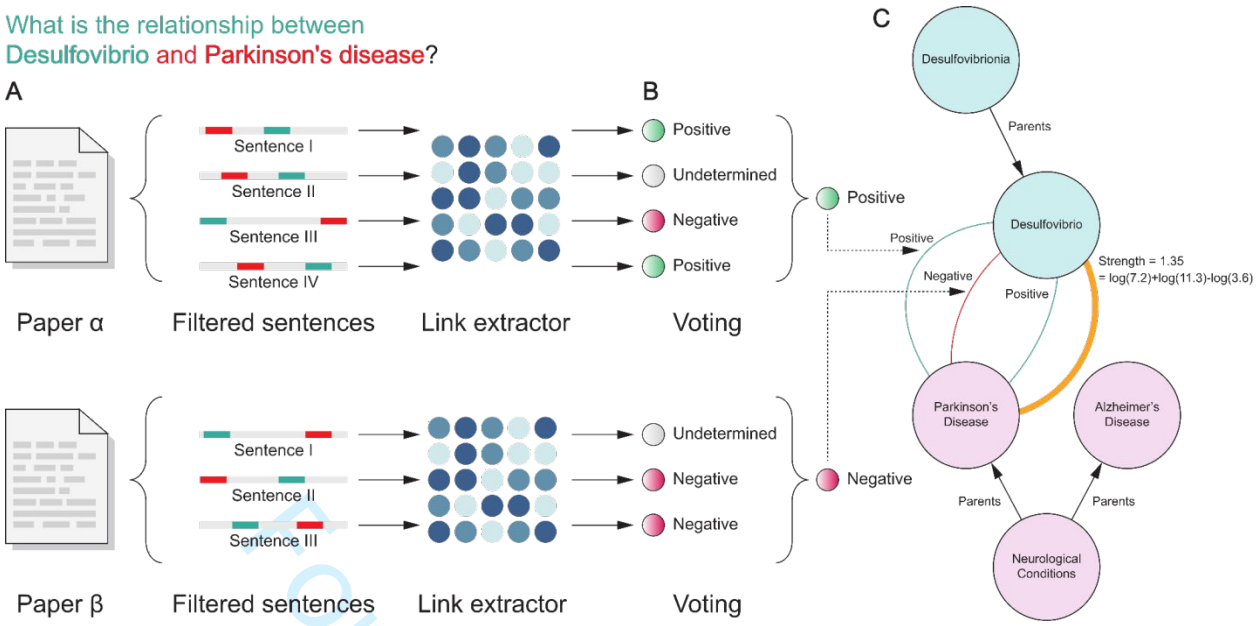

Fig. 6. Construction of MINERVA's robust knowledge graph. To add a microbe-disease relationship to our knowledge graph, a three-step process is followed. (A) For the candidate sentences, we just consider relationships in which our relation-extraction LLM is completely confident of their label. In practice, this is done by prompting our LLM five times with different temperatures and only incorporating the sentences in which the LLM always chooses the same label. (B) When multiple mentions of the same relationship occur within a publication, a majority vote determines the final label.(C) Finally, if the same relationship is mentioned across different publications, we incorporate the new information into our knowledge graph by updating the strength of the relationship using equation (2). Following this three-step approach, we ensure that all the information contained in MINERVA is highly accurate and reliable.

Supplementary Material:

MINERVA – Microbiome Network Research and Visualization Atlas:

A Scalable Knowledge Graph for Mapping Microbiome-Disease Associations

Authors: Saul Langarica[1,2], Young-Tak Kim[1], Adham Alkhadrawi[1], Jung Bin Kim[3], Synho Do[1,4,5*]

Affiliations:

[1] Laboratory of Medical Imaging and Computation, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, 125 Nashua Street, Boston, Massachusetts, USA.

[2] Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Macul, Santiago, Chile.

[3] Department of Neurology, Korea University Anam Hospital, Korea University College of Medicine, 73, Goryeodae-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea.

[4] KU-KIST Graduate School of Converging Science and Technology, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea.

[5] Kempner Institute, Harvard University, 150 Western Ave, Boston, Massachusetts, USA.

*Corresponding author.

A. LINK PREDICTION MODULE

To develop the *Link prediction* module, we implemented a graph neural network (GNN) model capable of predicting potential relationships between microbes and diseases not explicitly stated in the literature. We implemented a two-layer convolutional GNN architecture with ReLU activation functions as our graph encoder and a two-layer feed-forward neural network as the decoder, using PyTorch Geometric library [1].

The link prediction model was trained to classify potential links into three categories: No Link, Positive Relation, or Negative Relation. To optimize the model's performance, we experimented with different types of node features given as input to the model: (i) No Features: The GNN operated solely on the graph's structure, without any additional node information. (ii) Entity Definition Embeddings: We used a pre-trained sentence transformer [2] to generate embeddings of the textual definitions of each microbe and disease entity. (iii) Node2vec Embeddings: We employed Node2vec [3] to learn low-dimensional representations of nodes based on their structural roles in the graph. And (iv) Metapath2vec Embeddings: A similar technique to Node2vec but designed for heterogenous graphs [4].

To ensure a robust evaluation, we constructed our test set comprising 6,617 relationships where MINERVA's labels aligned with those of the benchmarked manually curated databases (see columns 3 and 4 in Table 1 in the main text). The remaining relationships were randomly split into training (80%) and validation sets (20%).

Table S1 presents the performance of the GNN on the test set, showcasing the results for each type of node embedding. Notably, both Node2vec and Metapath2vec embeddings, which capture the structural context of nodes within the graph, lead to the highest accuracy and F1 scores. However, the model without node embeddings follows closely, suggesting that the graph structure itself contains valuable information for link prediction, and node embeddings just offer a marginal performance improvement.

We ultimately selected the Node2vec embedding-based model as our final link prediction model due to its superior overall performance across the evaluated metrics. Future work will explore more sophisticated GNN architectures and the integration of additional node features, such as genomic data for microbes or disease ontology information, to further enhance link prediction accuracy.

Model	Accuracy	Precision	Recall	F1-Score
Metapath2vec	0.709	0.791	0.706	0.729
Node2vec	0.713	0.780	0.713	0.731
Sentence Embedding	0.660	0.660	0.808	0.696
No Embedding	0.702	0.778	0.699	0.725

Table S1: Link Prediction results for different types of node embeddings

B. ADDITIONAL EXAMPLES OF MINERVA EVIDENCE VS OTHER RESOURCES

Resource	Microbe	Disease	Resource Label	Resource Evidence	MINERVA Label	MINERVA Evidence
AMADIS	Bifidobacterium	Obesity	POSITIVE	PMID: 32309947	NEGATIVE	- **27845741**: Furthermore, an increase in the population of bifidobacterium exerts anti-obesity and lipid-lowering effects against high fat. - **35592636**: The use of probiotics containing lactobacillus and bifidobacterium species in obesity treatment is promising.
AMADIS	Escherichia Coli	Immflamation	NEGATIVE	PMID: 30897686	POSITIVE	- **37685055**: escherichia coli infection can disrupt this balance, increasing the abundance of gram-negative bacteria, increasing inflammation and oxidative damage, and disrupting the barrier function. - **34938203**: Interestingly, c. tropicalis was shown to positively correlate with serratia marcescens and escherichia coli in cd, further supporting their role in sustaining chronic inflammation as a “team” in the commensal niche (hoarau et al., 2016).
GMMAD	Lactobacillus	I. Bowel Disease	POSITIVE	PMID: 28039159	NEGATIVE	- **35215426**: Rosen et al. (2017) reported that inflammatory bowel disease was related to decreased abundance of microbes with anti inflammatory potential (such as bifidobacterium and lactobacillus). - **34641619**: In another trial, after patients with inflammatory bowel disease (inflammatory bowel disease) consumed yogurt, probiotics such as bifidobacterium and lactobacillus in the patients’ intestines increased, which helped improve intestinal function
GMMAD	Prevotella	Arthritis	NEGATIVE	PMID:	POSITIVE	- **30510245**: Thus, prevotella species are primary suspects

		Rheumatoid		18528968		also in humans, in which the increased abundance of these bacteria at mucosal sites has been associated with th17 mediated diseases including periodontitis 24 and rheumatoid arthritis 48 - **36933668**: In fact, numerous studies have identified an association between a disproportionate abundance of members of the prevotella taxa and a range of infections and inflammatory conditions including rheumatoid arthritis, intestinal and vaginal dysbiosis, metabolic disorders and major depressive disorder.
HMDAD	Collinsella aerofaciens	Colon Cancer	POSITIVE	PMID: 7574628	NEGATIVE	- **33615992**: Moreover, collinsella aerofaciens has been associated with a low risk of colon cancer, and patients with ibd show lower gut levels of this genus than do control individuals. - **32882999**: Collinsella aerofaciens has been associated with a low risk of colon cancer and ibs.
HMDAD	Bilophila	Liver cirrhosis	NEGATIVE	PMID: 25079328	POSITIVE	- **36957974**: Patients with cirrhosis showed higher abundance of enterococcaceae, gemellaceae at family level and phascolarctobacterium, enterococcus, streptococcus, gemella and bilophila at genus level in patients with hcc - **35897739**: Several in depth studies of patients with hcc have established associations between particular microbiome compositions and the development of hcc. For example, genera, such as bacteroides, phascolarctobacterium, enterococcus, streptococcus, gemella, bilophila, are increased in patients with hcc from nafld and cirrhosis compared to healthy controls.
DISBIOME	Streptococcus	Pneumonia	NEGATIVE	DOI:10.1017-S00071145 - 19001909	POSITIVE	- **35036248**: There are some studies that have reported the increased abundances of lachnospiracea e and ruminococcaceae in subjects with an insulin resistant status, such as diabetic disease and obesity. - **37891977**: While grape pomace promotes the decrease of enterobacteriacae and escherichia coli [251], a combination of quercetin and resveratrol leads to a reduced relative abundance of desulfovibrionaceae, acidaminococcaceae, coriobacteriaceae, bilophila, and lachnospiraceae (all possibly linked to diet induced obesity).
DISBIOME	Lactobacillus	Hypertension	POSITIVE	DOI:10.3389-fphar.2020 - 00258	NEGATIVE	- **36674891**: Lactobacillus is also able to help ameliorate hypertension by secreting substances working as th 17 lymphocytes inhibitors that decrease inflammation - **33212807**: The authors observed significantly lower association between the presence of lactobacillus in the group of patients with diabetes and hypertension, and higher in the diabetes only, and diabetes with hyperlipidemia cohorts.
MDIDB	Bifidobacterium	Melanoma	RELATE	PMID[1]: 31555274	NEGATIVE	- **29872574**: Matson et al. 10 collected 38 stool samples from melanoma patients on anti pd 1 treatment and after 16s RNA sequencing and quantitative PCR analysis he identified bifidobacterium spp, lactobacillus animalis, roseburia intestinalis and veillonella parvula as bacteria associated with beneficial response. - **33062956**: The presence of bifidobacterium in combination with anti pd l1 treatment can result in almost complete inhibition of melanoma tumor growth.
MDIDB	Helicobacter pylori	Ulcer	NEGATIVE	PMID[2]: 29576949	POSITIVE	- **34880265**: Perhaps the best known association is of bacteria (helicobacter pylori) causing gastric ulcers that progress into gastric cancer - **37600949**: helicobacter pylori (helicobacter pylori) is a bacterium that can live in the stomach and has been linked to many digestive disorders, including gastritis, stomach ulcers and stomach cancer.

Table S2: Additional examples of discrepancies between MINERVA and other resources

[1] In melanoma patients responding to iCPI more abundant species included Bifidobacterium, Collinsella, Enterococcus, Clostridiales, Rominococcus and Faecalibacterium, while low levels of Akkermansia muciniphila were observed in epithelial cancers not responding to iCPI (174).

[2] It is known that Helicobacter pylori plays an important role in the pathogenesis of ulcer, while Firmicutes are involved in obesity (Yu et al., 2014).

C. INDIVIDUAL RISK ASSESSMENT MODULE

Here the process for analyzing personal microbiome compositional data using the MINERVA platform is described. In this example, data from an Alzheimer's disease individual (SRR8061715) obtained from the Data Repository for Human Gut Microbiota's Project PRJNA496408 was analyzed (Experiment type: Amplicon; Instrument model: Illumina MiSeq; Geolocation: China). The microbiome abundance data for this patient is shown in Table S3.

NCBI Taxon ID	Relative Abundance	Scientific Name	Normalized Abundance (%)
1678	12.7199	Bifidobacterium	7.563911931
102106	0.314502	Collinsella	0.187019193
84111	0.244613	Eggerthella	0.14545957
133925	0.885265	Olsenella	0.526424461
838	2.21316	Prevotella	1.316059665
1253	0.232964	Pediococcus	0.138532471
1301	0.430984	Streptococcus	0.256285428
189330	0.768783	Dorea	0.457158225
841	0.675597	Roseburia	0.401744999
216851	2.6092	Faecalibacterium	1.551565579
204475	0.966803	Gemmiger	0.574911182
970	0.570763	Selenomonas	0.339405268
-1	136.331	Unknown	81.06947991
816	0.104834	Bacteroides	0.062339731
1263	5.75422	Ruminococcus	3.421757507
1485	3.34304	Clostridium	1.987944885

Table S3: Microbiome abundance data for the Alzheimer's disease patient (SRR8061715)

To begin, users must log in to the MINERVA system (Figure S1). Login options include direct account creation or Google account authentication. Without logging in, access to MINERVA is restricted.

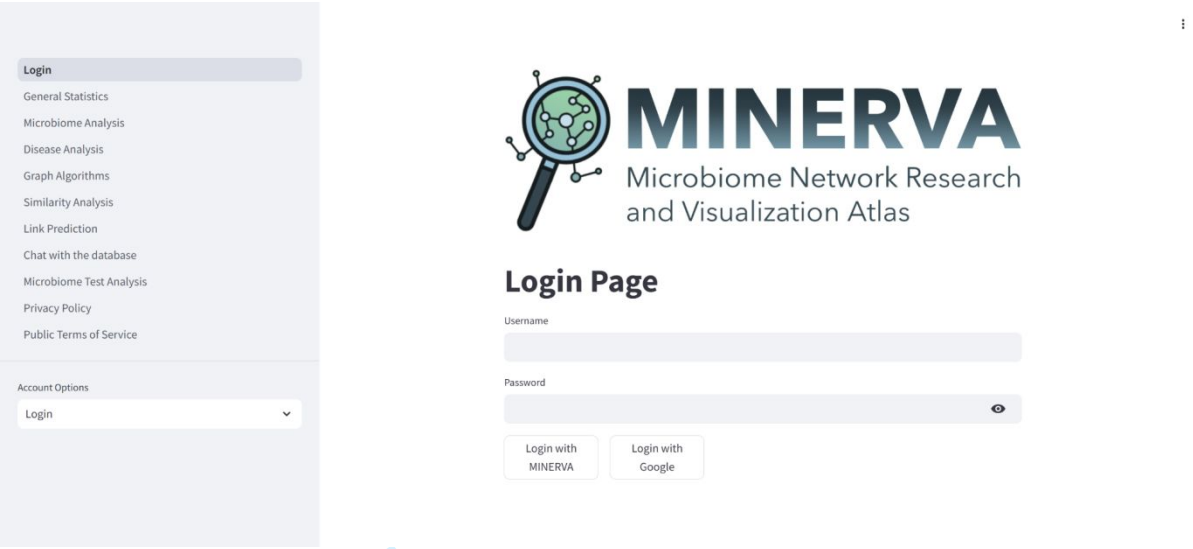

Figure S1: Login interface of the MINERVA system.

After logging in, users can select the *Microbiome Test Analysis* option in the left-side panel and input their microbial composition (Figure S2). Prior to entering individual microbiome data, demographic information for the healthy reference population must be specified. Selecting *Select all* displays the microbiome distributions for the entire healthy population using boxplots.

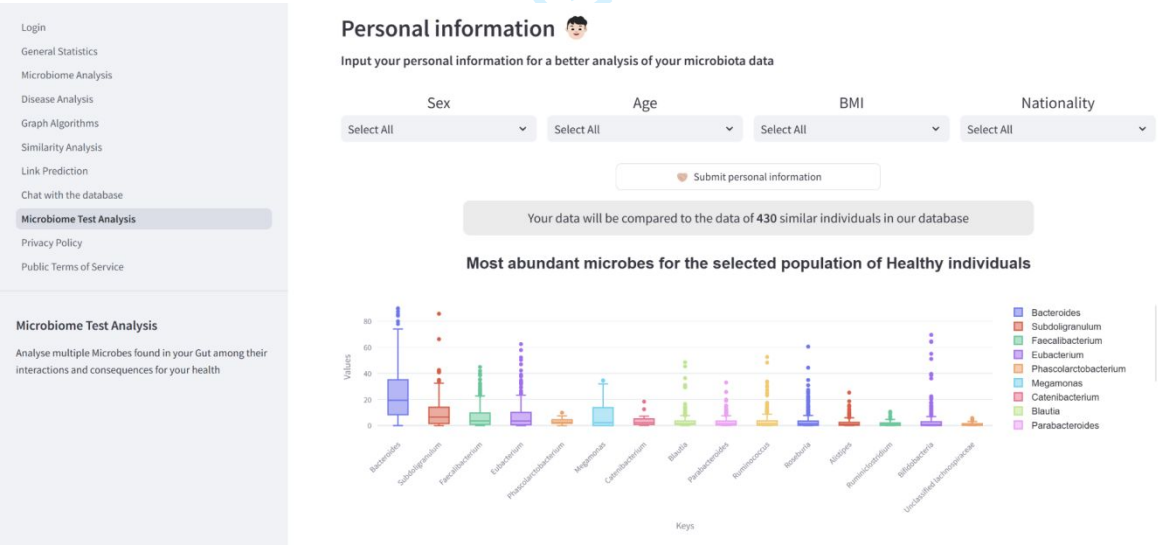

Figure S2: Personal information input interface and selection of healthy reference demographics.

Next, the *Add Microbe* button allows users to search for microbiome names and input relative abundances from personal samples. For this study, microbiome data from Table S3 was manually entered into the system. Alternatively, users can prepare a CSV file containing personal microbiome abundance data structured with two columns: *ncbi_taxon_id* and *relative_abundance*. By uploading this file to MINERVA, information can be entered easily without the need for manual input.

Upon clicking the *Analyze* button, the system shows the personal microbiome abundances relative to the interquartile range (IQR) of the healthy reference data (Figure S3). Values within the IQR are classified as normal, while those outside are flagged as too high or too low. For the Alzheimer's disease patient (SRR8061715), the abundances of Olsenella, Gemmiger, Selenomonas, and Paraclostridium exceeded the normal range, falling above the 95th percentile, whereas Bacteroides was below the normal range, falling below the 5th percentile.

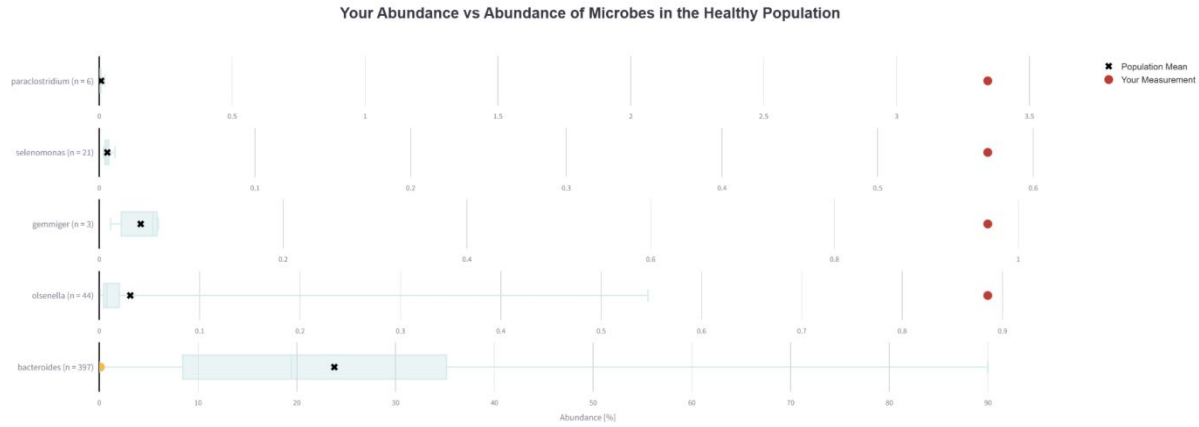

Figure S3: Position of personal microbiome abundances relative to the healthy reference IQRs.

The disease risks identified by MINERVA were subsequently summarized (Figure S4). A notable observation was the high risk of obesity, which aligns with previous studies [5, 6, 7]. Conversely, the patient exhibited a low risk for malignant neoplasms, consistent with reports indicating a reduced likelihood of cancer development in Alzheimer's disease patients [8, 9, 10]. Traditionally, gaining such insights from human personal microbiome abundance data would necessitate extensive literature reviews. MINERVA simplifies and accelerates this process, providing an efficient platform for in-depth analysis. Additionally, using an LLM, MINERVA generates detailed summary reports with references to further support personalized microbiome analysis (Figure S5).

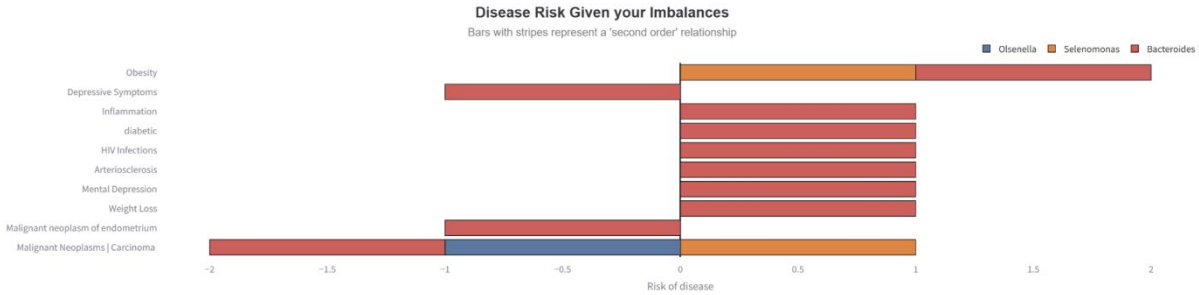

Figure S4: Disease risk summary from MINERVA for the Alzheimer's disease patient (SRR8061715).

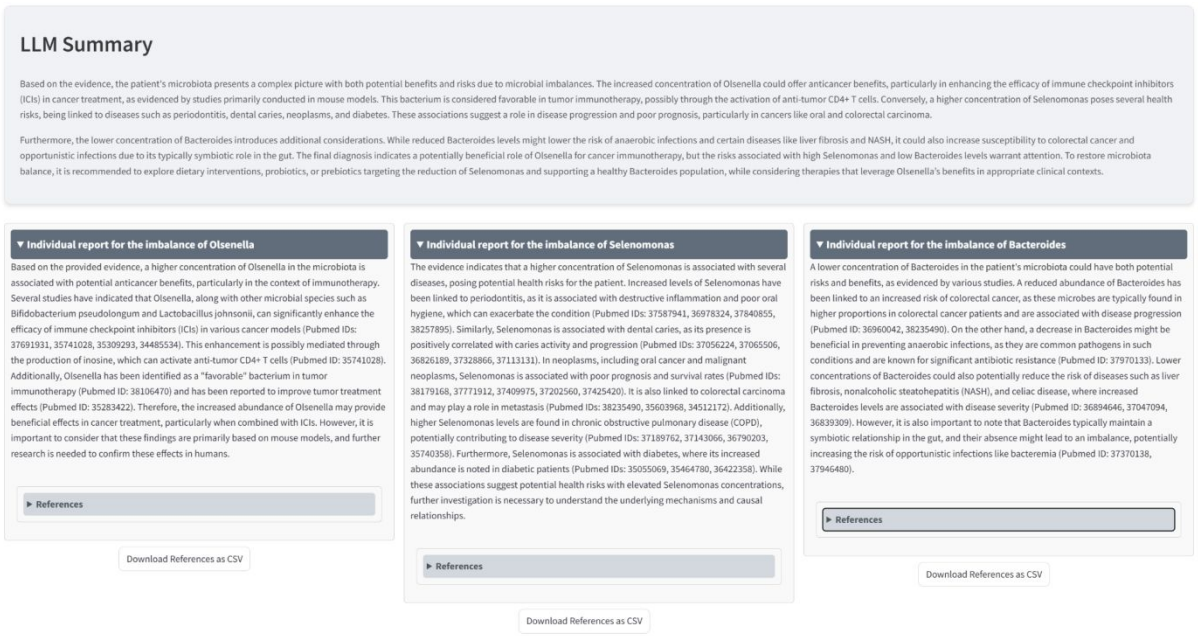

Figure S5: Example of a personal microbiome analysis report generated with a large language model.

D. ADDITIONAL RESULTS FOR THE POPULATION RISK ASSESSMENT MODULE

This study explores the outcomes of a case study utilizing the MINERVA system to analyze the microbiome composition of populations. Specifically, data were drawn from the Data Repository for Human Gut Microbiota's Project PRJNA496408, which involved Amplicon sequencing performed using the Illumina MiSeq platform, with samples sourced from a geolocation in China. The dataset included 32 individuals with amnestic mild cognitive impairment (MCI) and 33 individuals diagnosed with Alzheimer's disease (AD). It is important to note that the data and associated publication [11] used in this study were independent of the datasets utilized during the development of the MINERVA system.

The objective of the analysis was to identify disease-specific microbial imbalances (at the genus level) that could elucidate the role of microbiota in Alzheimer's disease progression and cognitive impairment risks at a group level. This analysis provided valuable insights into the evolving microbial signatures associated with disease development.

In its current configuration, the MINERVA system allows the analysis of a single target condition at a time. Accordingly, two separate analyses were conducted: one comparing Alzheimer's disease patients against a healthy reference population and another comparing individuals with impaired cognition to the same healthy reference population. This study presents and contrasts the insights gained from both analyses.

The Venn diagram analysis (Figure S6) reveals comparable numbers of differentially abundant microbes unique to each disease state when compared to the healthy population.

Notably, the genus *Hespellia* emerged as the sole taxonomic group present in both the cognitive impairment and Alzheimer's disease cohorts while being absent in the healthy reference population. Although, to the authors' knowledge, direct associations between *Hespellia* and these specific neurological conditions have not been previously documented in the literature, although there is some evidence that suggests a potential link between this genus and another neurological condition such as Parkinson's disease [12]. The consistent presence of *Hespellia* across both cognitive impairment and Alzheimer's disease warrants further investigation to elucidate its potential role in the progression of cognitive decline and its possible mechanistic involvement in neurodegenerative processes.

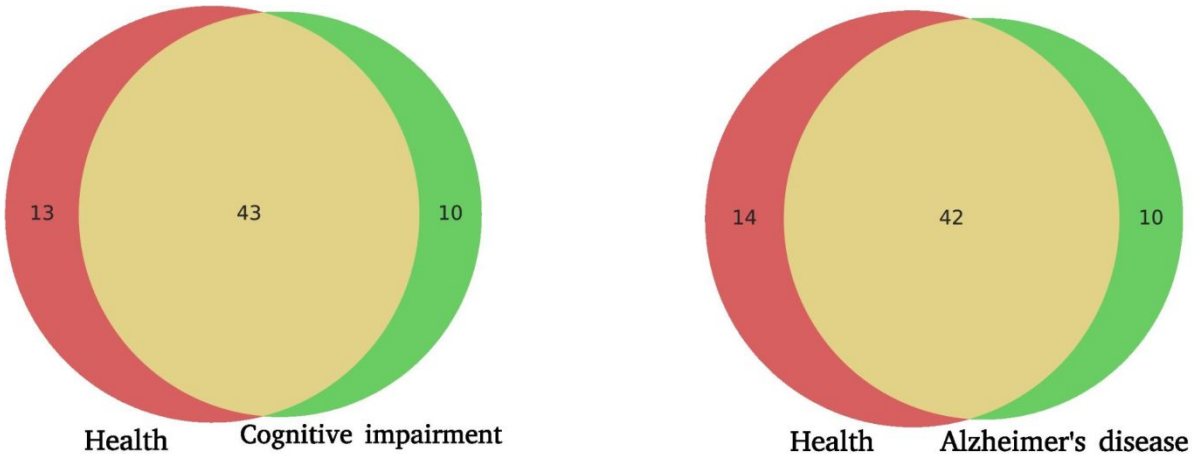

Figure S6: Venn Diagram generated by MINERVA to compare Genus present in the target and control populations.

Analysis of microbial diversity using MINERVA revealed both α and β-diversity patterns across the study populations. The α-diversity metrics (Figure S7) showed largely comparable diversity levels between disease states and the control group, with the notable exception of Simpson's Index, which demonstrated significant differences between the Alzheimer's disease and healthy populations. Assessment of β-diversity through PERMANOVA analysis of Bray-Curtis dissimilarity matrices indicated no significant differences in community composition between the studied populations.

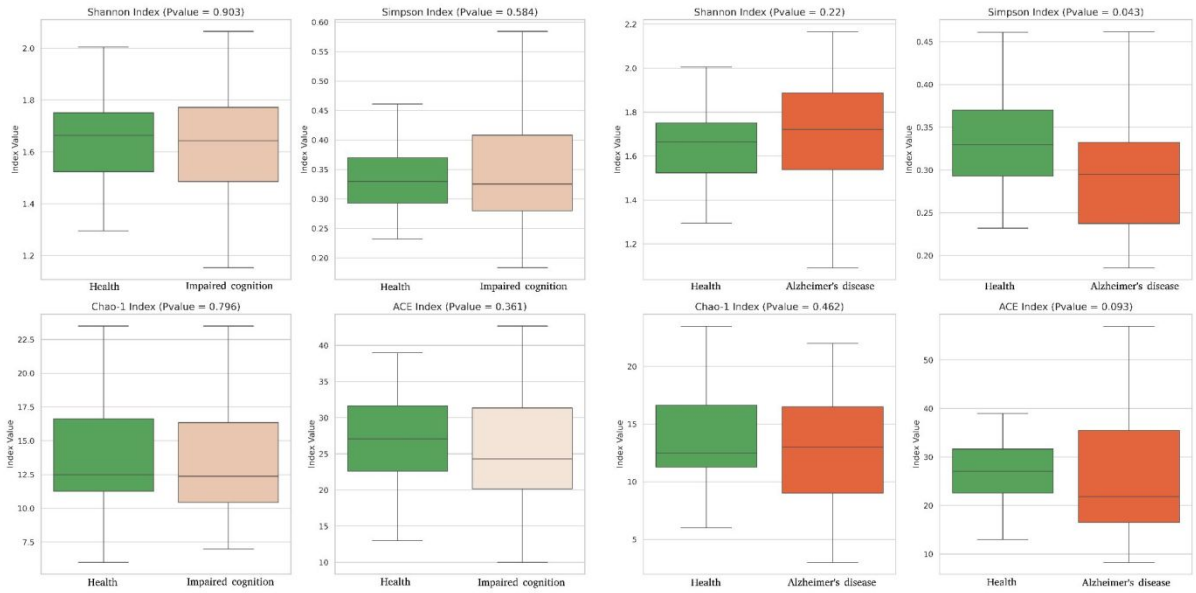

Figure S7: Different α-diversity metrics calculated by MINERVA comparing the target population with the control group.

Group-level analysis using Variable Importance in Projection (VIP) scores derived from Partial Least Squares-Discriminant Analysis (PLS-DA) (Figure S8) revealed distinct microbial signatures for each condition. The comparison between cognitive impairment and the healthy reference population identified *Ruminococcus*, *Bacteroides*, and *Defluviitalea* as the most discriminative genera. Conversely, *Bacteroides*, *Paraprevotella*, and *Defluviitalea* emerged as the key discriminative genera between Alzheimer's disease and healthy reference populations. The condition-specific importance of *Ruminococcus* in cognitive impairment and *Paraprevotella* in Alzheimer's disease suggests these genera may serve as potential biomarkers in the progression of cognitive decline.

Now, at the individual level, Figure S9 presents the distribution of microbiome abundances associated with both conditions. Microbial signatures linked to increased risk, marked in red, were primarily driven by *Bacteroides* in both MCI and AD patients. However, differences emerged between the two groups: An abnormal abundance of *Butyricicoccus* was identified as a contributing factor to AD risk [13] in patients with MCI, while abnormal abundances of Roseburia and Ruminococcus were identified as additional contributors to AD risk [14, 15], particularly in AD patients. This shift highlights the importance of stage-specific microbial profiling in understanding and mitigating Alzheimer's disease risk.

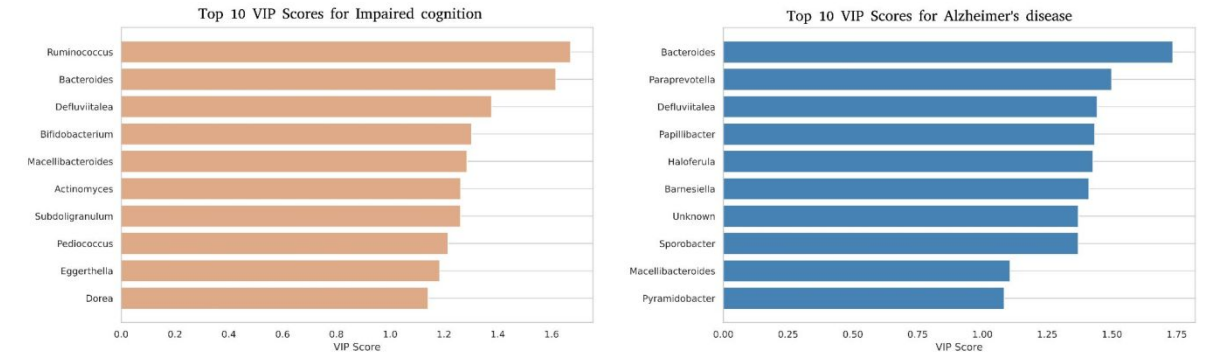

Figure S8: Variable importance in projection scores to show the most discriminative Microbes between the target and control populations, as determined by Partial Least Squares-Discriminant Analysis (PLS-DA).

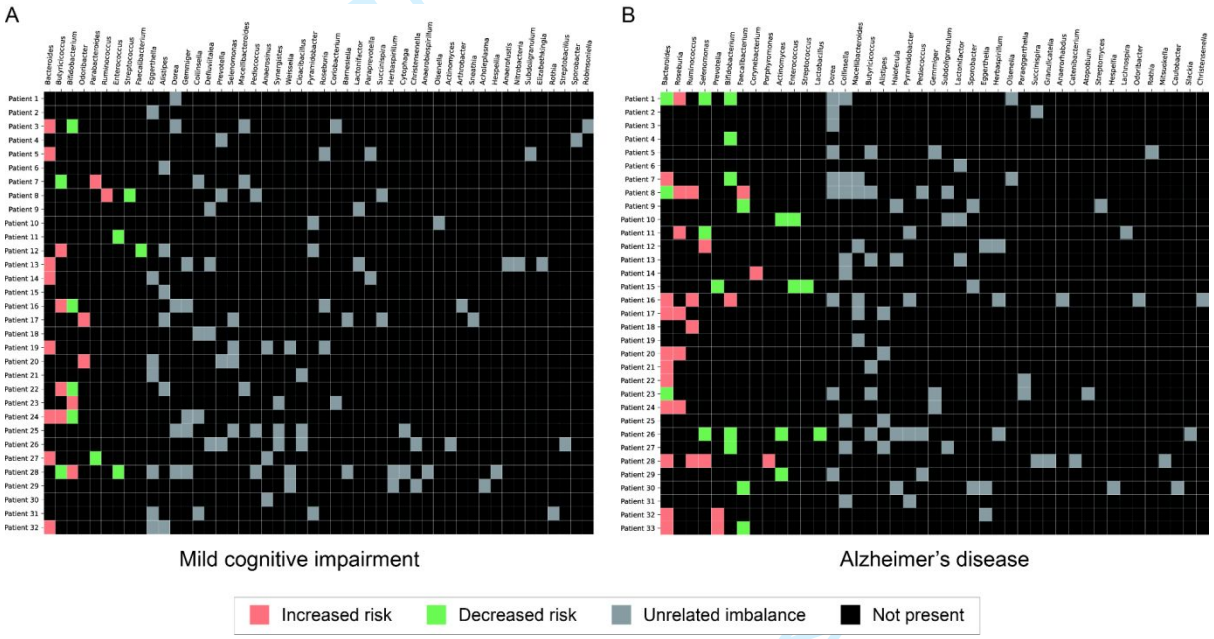

Figure S9: Alzheimer's disease risk induced by patient-level microbiome imbalances in MCI and AD. Heatmaps display Alzheimer's disease risk induced by abnormal microbiome imbalances. (A) MCI patients and (B) AD patients are shown. Red boxes represent risk-increasing microbes, green boxes indicate risk-decreasing microbes, gray boxes denote unrelated imbalances, and black areas show the absence of corresponding microbes abundance-induced risk association.

The relative rankings of disease-associated risks in the two patient groups are illustrated in Figure S10. Emotional and fatigue-related conditions, such as anxiety, depressive symptoms, and fatigue, ranked highly in MCI patients, reflecting their close association with early disease stages. By contrast, systemic and metabolic conditions, including osteoporosis, weight gain, and dysbiosis, were prominent in the AD group. Notably, Alzheimer's disease itself exhibited consistently high rankings, accompanied by a broader spread of risk-related conditions compared to MCI. This increasing heterogeneity in AD patients suggests a

diversification of comorbidities as the disease progresses, pointing to the cumulative effects of long-term microbial dysregulation.

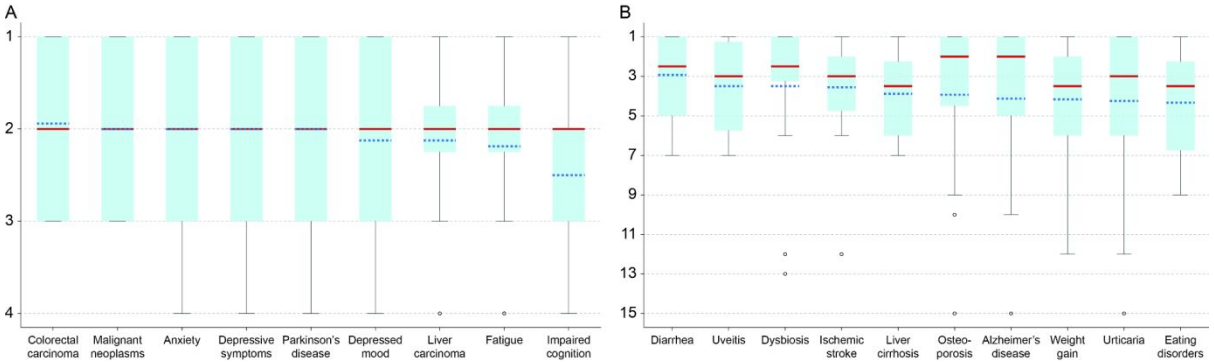

Figure S10: Ranking of disease-associated risks in MCI and AD patients}. Boxplots show the ranking of disease risks in (A) MCI and (B) AD patients. Red lines indicate medians and blue dashed lines represent means.

Overall, the findings from this small-scale case study demonstrate the utility of MINERVA for population microbiome analysis. Distinct microbial imbalances and their associated disease risks were observed across both group and individual levels, providing new insights into the evolving gut microbiome contributions to Alzheimer's disease pathogenesis.

References

1. Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric. In: ICLR2019 Workshop on Representation Learning on Graphs and Manifolds; 2019.
2. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates Inc..
3. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 855–864.
4. Dong Y, Chawla NV, Swami A. metapath2vec: Scalable Representation Learning for Heterogeneous Networks.In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 135–144.
5. Terzo S, Amato A, Mul`e F. From obesity to Alzheimer's disease through insulin resistance. J Diabetes Complications. 2021 Aug;35(11):108026.
6. Hinney A, Albayrak O, Antel J, Volckmar AL, Sims R, Chapman J, et al. Genetic variation at the CELF1 (CUGBP, elav-like family member 1 gene) locus is genome-wide associated with Alzheimer's disease and obesity. Am J Med Genet B Neuropsychiatr Genet. 2014 May;165B(4):283-93.

7. Litwiniuk A, Bik W, Kalisz M, Baranowska-Bik A. Inflammasome NLRP3 Potentially Links Obesity-Associated Low-Grade Systemic Inflammation and Insulin Resistance with Alzheimer's Disease. Int J Mol Sci. 20212 May;22(11).
8. Shi Hb, Tang B, Liu YW, Wang XF, Chen GJ. Alzheimer disease and cancer risk: a meta-analysis. Journal of Cancer Research and Clinical Oncology. 2015 Mar;141(3):485-94.
9. Ou SM, Lee YJ, Hu YW, Liu CJ, Chen TJ, Fuh JL, et al. Does Alzheimer's disease protect against cancers? A nationwide population-based study. Neuroepidemiology. 2012 Oct;40(1):42-9.
10. Musicco M, Adorni F, Di Santo S, Prinelli F, Pettenati C, Caltagirone C, et al. Inverse occurrence of cancer and Alzheimer disease: a population-based incidence study. Neurology. 2013 Jul;81(4):322-8.
11. Liu P, Wu L, Peng G, Han Y, Tang R, Ge J, et al. Altered microbiomes distinguish Alzheimer's disease from amnestic mild cognitive impairment and health in a Chinese cohort. Brain, Behavior, and Immunity. 2019;80:633-43.
12. Hey G, Nair N, Klann E, Gurrala A, Safarpour D, Mai V, et al. Therapies for Parkinson's disease and the gut microbiome: evidence for bidirectional connection. Front Aging Neurosci. 2023 May;15:1151850.
13. Nguyen TTT, Fujimura Y, Mimura I, Fujii Y, Nguyen NL, Arakawa K, et al. Cultivable butyrate-producing bacteria of elderly Japanese diagnosed with Alzheimer's disease. J Microbiol. 2018 Aug;56(10):760-71.
14. Li H, Cui X, Lin Y, Huang F, Tian A, Zhang R. Gut microbiota changes in patients with Alzheimer's disease spectrum based on 16S rRNA sequencing: a systematic review and meta-analysis. Front Aging Neurosci. 2024 Aug;16:1422350.
15. Heravi FS, Naseri K, Hu H. Gut Microbiota Composition in Patients with Neurodegenerative Disorders (Parkinson's and Alzheimer's) and Healthy Controls: A Systematic Review. Nutrients. 2023 Oct;15(20).

D. PROMPT FOR RELATION EXTRACTION

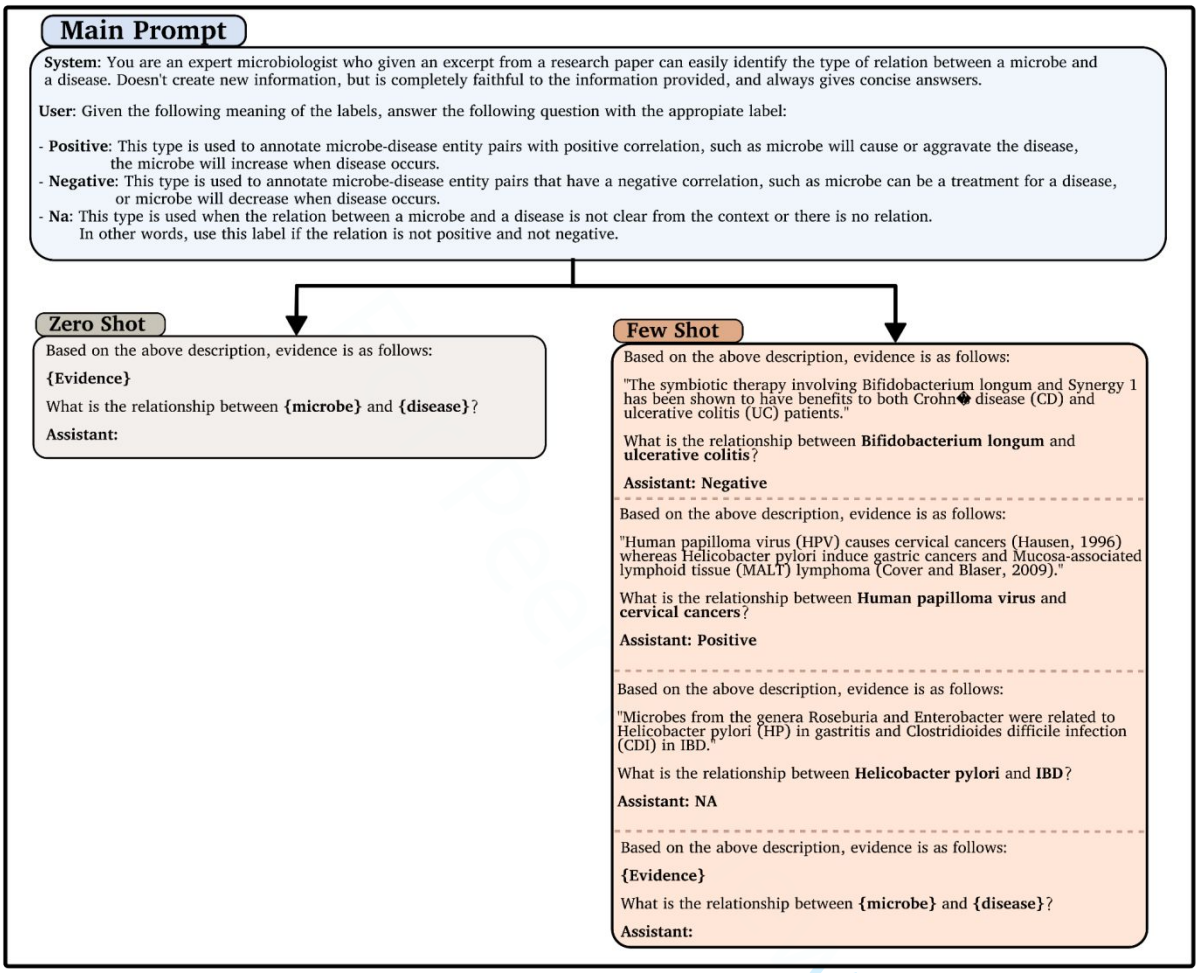

Figure S11: Prompt for relation-extraction models}. Prompt templates for zero-shot (left) and few-shot (right) models. Fine-tuned models utilized the zero-shot prompt template.