TU WIEN Informatics

Record Linkage über Metagenomische Microbiome Profile

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Business Informatics

eingereicht von

Li I Wu
Matrikelnummer 12320187

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber
Mitwirkung: Univ.Lektor Dipl.-Ing. Mag.rer.soc.oec. Rudolf Mayer

Wien, 31. August 2026

Li I Wu	Andreas Rauber

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Wien Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Informatics

Record Linkage Using Metagenomic Microbiome Profiles

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieurin

in

Business Informatics

by

Li I Wu
Registration Number 12320187

to the Faculty of Informatics

at the TU Wien

Advisor: Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Andreas Rauber
Assistance: Univ.Lektor Dipl.-Ing. Mag.rer.soc.oec. Rudolf Mayer

Vienna, August 31, 2026

Li I Wu | Andreas Rauber

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Li I Wu

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, habe ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT-Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 31. August 2026

Li I Wu

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

I would like to express my deepest appreciation to Rudolf Mayer for guiding me throughout the entire process of this thesis and providing insightful advice. I would also like to thank Andreas Rauber for helping to refine the focus of this thesis and providing valuable feedback.

I would like to express my sincere gratitude to my mother, Siou-Mei Chen, for her unconditional love and support. I would also like to thank my father, Kuo-Wei Wu, for everything he has done for me, and Patrick Styll for his steadfast presence and engaging discussions about this thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

Personal microbiome identification (PMI) links individuals by matching microbiome samples collected at different points in time, providing a means of record linkage. Among previous works, three PMI approaches include the metagenomic-code method, GePMI, and the nearest-neighbour method. These approaches differ in their feature abundance thresholds, feature representations, and matching mechanisms. Although the studies introducing these approaches demonstrated their feasibility, subsequent comparisons have included only subsets of them. No study has evaluated all proposed PMI approaches within a unified setting, leaving their relative performance unclear.

In this thesis, we evaluated all proposed PMI approaches within a unified setting that more closely reflects real-world conditions than previous studies. In terms of the samples included, the evaluation setting involved a more challenging study population, greater longitudinal diversity, and a higher number of true negatives. This allowed us to identify the matching scenarios in which the different PMI approaches perform best and inform their future application.

A further focus of this thesis is the association between linkage errors and PMI approaches. At the sample level, we analysed sample interactions and method-related factors and used the findings to identify strategies to improve PMI performance. The mutual nearest-neighbour (MNN) method without rematching increased the average F1 score by approximately 2% compared with the nearest-neighbour method in evaluation scenarios including true negatives. At the feature level, we analysed feature contribution patterns and identified the features that frequently dominated each linkage outcome.

As a final focus, given the privacy concerns arising from the unintended or unauthorised use of PMI, we evaluated privacy-preserving techniques, including data synthetisation and data anonymisation, as safeguards for microbiome data shared for research purposes.

Taken together, this thesis provides a comparative evaluation of PMI approaches, investigates patterns underlying linkage errors, and evaluates privacy-preserving techniques for microbiome data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Abstract **ix**

Contents **xi**

1 Introduction **1**
1.1 Problem Statement . . . 1
1.2 Research Questions . . . 2
1.3 Thesis Structure . . . 4

2 Background **5**
2.1 Microbiome Data . . . 5
2.2 Microbiome collection and analysis workflow . . . 6
2.3 Types of raw microbiome data . . . 7

3 State of the Art **9**
3.1 Personal Microbiome Identification . . . 9
3.2 Privacy-preserving techniques . . . 17

4 Dataset **19**
4.1 Data Understanding . . . 19
4.2 Data Acquisition and Bioinformatics Pipeline . . . 23

5 Experiment Design **27**
5.1 Personal Microbiome Identification . . . 27
5.2 Linkage Errors . . . 34
5.3 Privacy-Preserving Techniques . . . 41

6 Evaluation **47**
6.1 Personal microbiome Identification . . . 47
6.2 Linkage Errors . . . 53
6.3 Privacy-Preserving Techniques . . . 64

7 Conclusion **69**
7.1 Research Questions . . . 69

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

7.2 Future work 71

Overview of Generative AI Tools Used **73**

List of Figures **75**

List of Tables **77**

List of Algorithms **81**

Appendix **83**

Bibliography **89**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 1

Introduction

1.1 Problem Statement

The human body is an ecosystem where trillions of tiny organisms coexist with the host. The human microbiota refers to the collection of microorganisms, such as bacteria, archaea, fungi, and viruses, that inhabit nearly all parts of the human body, and the human microbiome refers to the set of genes of these microorganisms [AKP+23, UMPK12]. Microbiota composition reflects the host's genetic, demographic, lifestyle, and environmental factors and serves as an important indicator for disease prediction, diagnosis, and therapy. For example, colonisation by Helicobacter pylori in the stomach is related to gastric cancer [LSL+19]; a lower abundance of Firmicutes in the intestines is linked to inflammatory bowel disease [SPW+08, WDH+10, MJS+14]; and a lower abundance of Proteobacteria in the airway and lungs is connected to asthma [MSW+13]. Given these associations, microbiome data are increasingly collected in clinical and laboratory repositories worldwide for research purposes.

Record linkage is the task of identifying records that correspond to the same entity, such as individuals, companies, geographic regions, families, or households, from one or more data sources [GBVR03]. In healthcare, record linkage enables researchers to assess care integration, consider long-term patient outcomes, and monitor service provision across different sectors or agencies. Deterministic linkage is applied when one or several identifiers, such as a health care service number, may be used alone or in combination [BJS+10]. However, these identifiers are not always reliable, as linkage errors can arise from poor data quality, missing values, domain-specific identifiers that are not shared across sources, or encrypted identifiers employed for privacy protection [HDB+17]. Probabilistic linkage is typically applied when no unique identifier is available and involves matching partially identifying variables, such as name, address, age, and gender, which may not be unique [BJS+10, GBVR03].

Previous studies show that the gut microbiome is fairly stable over timescales of many months or years, with smaller intra-individual variation than inter-individual variation [SP19]. This raises the question of whether the microbiome can serve as a probabilistic linkage identifier, enabling records corresponding to the same individual to be identified based on their microbiota composition. Personal microbiome identification (PMI) refers to techniques that link individuals by matching microbiome samples collected at different points in time. Existing PMI approaches include the metagenomic-code method [FHM+15], the method called "Generating inter-individual similarity distribution for Personal Microbiome Identification" (GePMI) [WLW+18], and the nearest-neighbour method [MHE24]. The studies introducing these methods demonstrate the feasibility of PMI on small- to medium-sized datasets involving at most 155 sample pairs from a single body site.

Microbiome data are often shared across multiple parties for research purposes, which may increase the risk of unintended or unauthorised use of PMI, thereby raising privacy concerns. Combining information about an individual's microbiome data with their genetic and medical data has the potential to reveal highly personal details, such as disease susceptibility, travel history, alcohol consumption, and drug use [PDLDDT25]. Recently studied privacy-preserving techniques for microbiome data include data anonymisation using k-anonymity via Semantic Microaggregation [MKH23], as well as data synthetisation using Classification and Regression Trees (CART) with synthpop and a Gaussian copula model with the Synthetic Data Vault (SDV) [HME22]. Both studies show promising results in balancing privacy protection, statistical fidelity, and analytical utility.

Several research gaps exist in the field of PMI. First, the three PMI approaches mentioned above have not been evaluated jointly within a unified setting, leaving their relative performance unclear. Second, real-world record linkage scenarios are typically more complex than those considered in previous studies, and it remains unclear whether PMI approaches maintain their performance in such settings. Accordingly, the main contribution of this thesis is a comparative evaluation of PMI approaches in settings more challenging than those considered in previous studies. In addition, we investigate patterns associated with linkage errors to identify potential improvements to PMI approaches. Finally, we evaluate privacy-preserving techniques on microbiome data in terms of their ability to safeguard privacy while preserving statistical fidelity.

1.2 Research Questions

1. **To what extent are personal microbiome identification (PMI) approaches successful as record linkage techniques, based on metrics such as precision, recall, and F1-score, when applied in more challenging PMI settings than those considered in previous studies, including:**
   a) **A more homogeneous study population,**
   b) **Greater longitudinal diversity, and**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

c) **A higher number of true negatives?**

In this research question, we apply the PMI approaches, namely the metagenomic-code method, GePMI, and the nearest-neighbour method, to a selected microbiome dataset involving samples collected from participants at different time points. Each microbiome sample collected from a specific individual at a specific time point represents a record, and each individual represents the linkage entity. Linkage performance is evaluated using precision, recall, and F1-score.

Each of the three subquestions, RQ1.a), RQ1.b), and RQ1.c), represents a characteristic that makes PMI settings more challenging than those considered in previous studies and forms the basis of the dataset selection criteria.

For RQ1.a), we apply the PMI approaches to datasets larger than those used in previous studies. The selected dataset includes more than 250 participants, each providing samples at two time points, to examine how an increased sample size affects linkage performance. For RQ1.b), we apply the PMI approaches to datasets with more time points and more diverse intervals between sample collections than those used in previous studies. The selected dataset includes more than 50 participants, each providing samples at at least three time points with short- and intermediate-term intervals, to examine how variation in collection intervals affects linkage performance. For RQ1.c), we use the datasets from RQ1.a) and RQ1.b) as baselines and apply PMI techniques to datasets with a higher number of records from different individuals to examine how increased true-negative cases affect linkage performance.

2. **Which patterns associated with linkage errors can be identified as potentially useful for guiding improvements to PMI approaches in terms of:**

   a) **Sample interactions, and**

   b) **Feature contributions?**

In this question, we focus on cases with linkage errors, identify patterns of these errors and the features associated with them, and investigate whether these patterns can be mapped to possible interventions, thereby providing directions for improving the performance of PMI approaches as record-linkage techniques.

The two subquestions, RQ2.a) and RQ2.b), represent different targets of analysis. For RQ2.a), we conduct feature importance analyses to identify which taxa are associated with linkage errors. For the metagenomic-code method, which relies on the hitting-set mechanism, we analyse feature-level contributions to false-positive and false-negative cases. For GePMI and the nearest-neighbour method, which rely on distance and similarity measures, we analyse feature-level contributions to the resulting distances and similarities based on different metrics, such as correlation distance and Jaccard similarity. For RQ2.b), we categorise potential causes of linkage errors, such as outlier bias, data quality issues, antibiotic use, probiotic use,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

and dietary factors, based on the features associated with linkage errors identified in RQ2.a).

3. **To what extent are data anonymisation and data synthetisation techniques successful as safeguards against personal microbiome identification (PMI) in terms of protecting microbiome data while maintaining statistical fidelity?**

   In this question, we apply synthetic data generation approaches, such as those implemented in the Synthetic Data Vault (SDV), to generate synthetic data, and micro-aggregation tools, such as (μ-ANT), to generate anonymised data. We assess their statistical fidelity using correlation and decomposition analyses, and evaluate their analytical utility through machine learning (ML) tasks such as disease prediction and species prediction, by training identical models on each dataset and comparing their ML performance using metrics such as auROC and auPRC.

1.3 Thesis Structure

The remainder of this thesis is organised as follows. Chapter 2 presents the background information. Chapter 3 presents the state of the art, covering PMI approaches and privacy-preserving techniques for data anonymisation and data synthetisation. Chapter 4 describes the dataset used in this thesis. Chapter 5 describes the experiment design for each research question in this thesis. Chapter 6 presents the evaluation for each experiment in this thesis. Finally, Chapter 7 summarises the main contributions of this thesis.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 2

Background

This chapter provides background information relevant to this thesis. Section 2.1 introduces the microbiome and feature abundance profiles as a common data representation in microbiome research. Section 2.2 outlines the process of deriving feature abundance profiles, and Section 2.3 describes the different types of raw microbiome data from which these feature abundance profiles can be derived.

2.1 Microbiome Data

The human microbiota refers to the collection of microorganisms, such as bacteria, archaea, fungi, and viruses, that inhabit nearly all parts of the human body. These microorganisms are found at both external and internal body sites. External sites include the skin, eyes, and areas beneath the nails, while internal sites include the lungs, gut, bladder, kidneys, and vagina. They are also found at portals of entry, including the respiratory tract (mouth and nose), gastrointestinal tract (oral cavity), urogenital tract, and breaks in the skin. Microbiota composition varies across body sites and is highly personalised, differing among individuals [AKP+23, UMPK12].

The term "human microbiome" has been defined in various ways. Some definitions refer to the set of genes of the microorganisms, while a broader definition includes the microorganisms themselves and their "theatre of activity", comprising structural elements, metabolites, molecules produced by coexisting hosts, and the surrounding environmental conditions [BRF+20].

Microbiome interactions play a key role in human health, including in the digestive, respiratory, urinary, and reproductive systems, as well as the skin [AKP+23]. For example, colonisation by Helicobacter pylori in the stomach is related to gastric cancer [LSL+19]; a lower abundance of Firmicutes in the intestines is linked to inflammatory bowel disease

[SPW+08, WDH+10, MJS+14]; and a lower abundance of Proteobacteria in the airway and lungs is connected to asthma [MSW+13].

In microbiome research, feature abundance profiles are a common data representation providing the abundance of each feature in a sample. Depending on the research objective, the features may represent different types of biological information. This thesis focuses on microbial taxa, which are groups of microorganisms classified at different taxonomic ranks, such as kingdom, phylum, class, order, family, genus, species, and strain. For example, at the species level, the profile shows the abundance of each microbial species in a microbiome sample.

2.2 Microbiome collection and analysis workflow

To derive and analyse feature abundance profiles in microbiome research, the complete process comprises sample collection, experimental procedures, a bioinformatics pipeline, and analysis, as described in detail below and illustrated in Figure 2.1 [KLFP24].

1. Sample collection (sampling): Microbiome samples originate from the host or environment. In the context of the human microbiome, samples may be collected from various body sites, including the skin, oral cavity, gut, and vagina [AKP+23].

2. Experimental procedure: Collected samples are processed using different experimental methods, such as shotgun sequencing, amplicon sequencing, metatranscriptomics, metabolomics, and metaproteomics, to generate raw microbiome data.

3. Bioinformatics pipeline: Raw microbiome data are processed using computational methods, including assembly-based profiling, reference-based classification, metabolite profiling, and protein profiling, to derive biologically meaningful outputs.

4. Analysis: The outputs of bioinformatics pipelines are used in analyses.

The methods and tools used in the experimental procedure and bioinformatics pipeline stages vary depending on the research objective. When reviewing previous studies, the representation of microbiome data as feature abundance profiles alone does not necessarily ensure comparable evaluations across studies. The type of raw microbiome data generated during the experimental procedure stage and the bioinformatics pipeline used to derive the feature abundance profiles, which are usually reported in the studies, must also be taken into account.

In this thesis, the study accompanying the selected dataset serves as the basis for deriving the feature abundance profile. We download raw microbiome data from a public repository and run the bioinformatics pipeline following the procedures described in the study. The resulting feature abundance profile is described in detail in Chapter 4.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

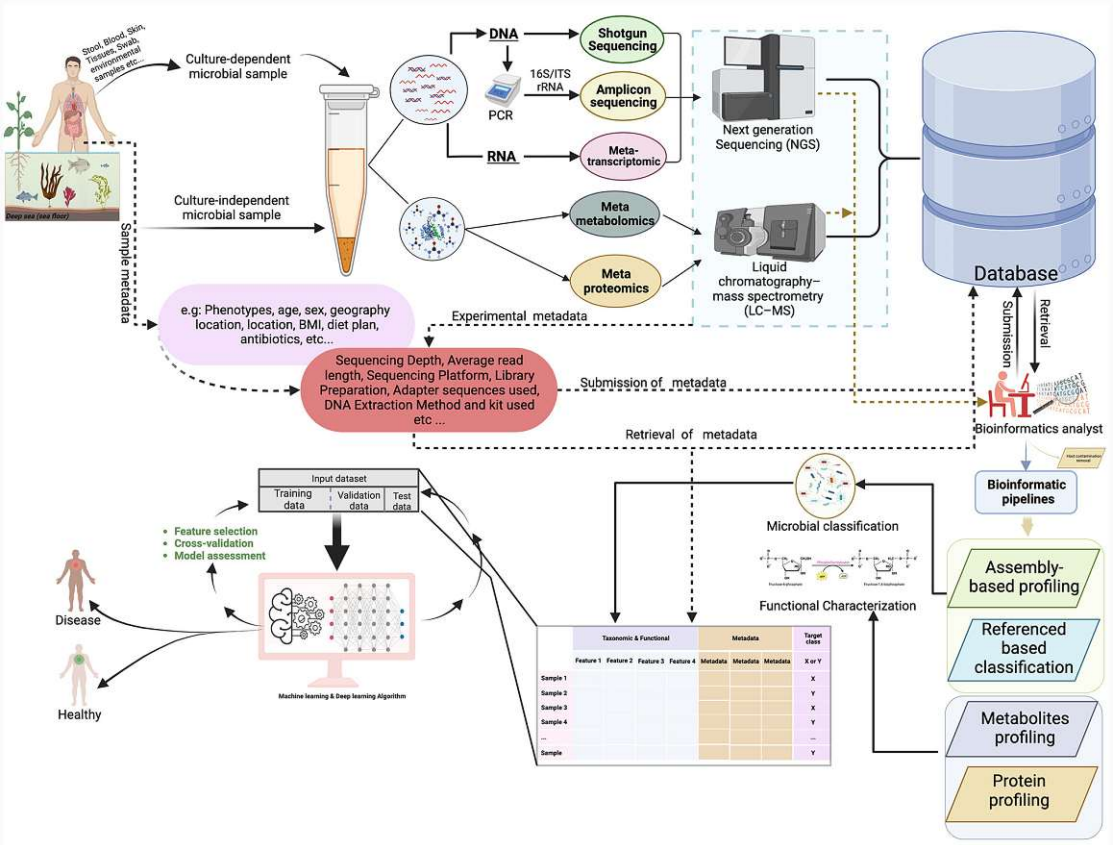


Figure 2.1: Microbiome collection and analysis workflow. The stages include (i) sample collection (top-left), (ii) experimental procedure (top-middle), (iii) bioinformatics pipeline (bottom-right), and (iv) analysis (bottom-left). Reproduced from [KLFP24], licensed under CC BY 4.0.

2.3 Types of raw microbiome data

When reviewing previous studies, the corresponding types of raw microbiome data are a key factor in determining whether evaluations are comparable across studies. Table 2.1 outlines common raw microbiome data types, which are typically defined by the methods used to generate them.

In the studies accompanying the PMI approaches, the Franzosa dataset is used to evaluate both the metagenomic-code method [FHM+15] and the nearest-neighbour method [MHE24]. The dataset includes four metagenomic feature types: OTUs, Species, Markers, and kbwindows. The raw microbiome data type for OTUs is amplicon sequencing of the 16S rRNA gene, and the raw microbiome data type for the other three metagenomic feature types is whole-metagenome shotgun sequencing (WMS).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 2.1: Raw microbiome data types. Adapted from [KLFP24], licensed under CC BY 4.0. Capabilities*: Resolution, the ability to distinguish between different microbes or genes; Selectivity, the ability to target specific microbes or genes for analysis; Capacity, the amount of data that can be generated and analysed.

Data type	Definition	Capabilities*
Shotgun metagenomics (Whole-metagenome shotgun sequencing, WMS)	Whole-genome sequencing of all genomes in a sample, including DNA from bacteria, fungi, viruses, and the host organism	High resolution Moderate selectivity High capacity
Amplicon sequencing	Targeted sequencing of a specific gene (e.g., 16S rRNA gene) or region of DNA from a sample	Low resolution High selectivity Medium capacity
Meta-transcriptomics	Whole-transcriptome sequencing of all RNA transcripts in a sample, including RNA from bacteria, fungi, viruses, and the host organism (e.g. a human person)	High resolution Moderate selectivity High capacity
Metabolomics	Identification and quantification of all metabolites in a sample	Low resolution Low selectivity High capacity
Metaproteomics	Study of the entire protein collection (proteome) of a microbial community	Low resolution

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 3

State of the Art

This chapter presents the state of the art in the topic of this thesis. Section 3.1 introduces personal microbiome identification and its approaches, including the metagenomic-code method, GePMI, and the nearest-neighbour method. Section 3.2 outlines privacy-preserving techniques, including data anonymisation and data synthetisation.

3.1 Personal Microbiome Identification

As a means of record linkage, personal microbiome identification (PMI) is defined here as techniques that link individuals by matching microbiome samples collected at different points in time.

A common use case involves longitudinal samples from one or more cohort studies, in which a set of samples is collected at an earlier time point, and another set is collected at a later time point. The samples from the later time point are then matched to those from the earlier time point. Previous studies do not use uniform terminology for the two sets of samples. [FHM+15] refer to them as first-visit and second-visit samples. However, PMI is not limited to this temporal order, and some datasets include more than two visits. [WLW+18] uses terms such as given, collected, or target samples and query, newly acquired, or test samples. In this thesis, they are referred to as *known samples* and *query samples*, respectively. PMI matches query samples to known samples when they are identified as originating from the same individual.

In the literature, [FHM+15] first proposed the metagenomic-code method and evaluated it with samples from the Human Microbiome Project (HMP) cohort. Subsequently, [WLW+18] proposed a method called GePMI, evaluated it with samples from multiple cohorts, and compared it with the metagenomic-code method on a subset of the dataset used in [FHM+15]. Finally, [MHE24] discussed both prior PMI approaches, proposed

the nearest-neighbour method, and evaluated it against the metagenomic-code method on the dataset used in [FHM+15].

3.1.1 Approaches for PMI

The metagenomic-code method

The metagenomic-code method [FHM+15] matches samples based on discriminative feature sets constructed from known samples, using detection and non-detection thresholds to define the presence and absence of features.

For each sample, threshold-based discretisation is applied to classify features according to their abundance as confidently detected, ambiguous, or confidently not detected (see Table 3.1).

Table 3.1: Detection thresholds in the metagenomic-code method. Adapted from [FHM+15]. Features with abundance between the confident detection and confident nondetection thresholds are considered ambiguous.

Feature level	Confident detection threshold	Relaxed detection threshold	Confident nondetection threshold
Taxon	$> 10^{-3}$	$> 10^{-4}$	$< 10^{-5}$
Gene	> 5	> 0.5	< 0.05

A metagenomic code is a set of discriminative features that distinguishes a known sample from other known samples. It is constructed using a hitting set-based greedy algorithm. For each known sample, features are ranked by abundance gap, that is, the difference between their abundance in the known sample and their next highest abundance among the other known samples. In this way, features that are more discriminative and less likely to be acquired by other known samples over time are prioritised.

The metagenomic code is constructed as follows. For known sample i, let F denote the vector of confidently detected features, ranked by descending abundance gap, S an initially empty set representing the metagenomic code, and J the set of all other known samples. Iteratively, the highest ranking feature f is removed from F. If f is not present in the feature vector of one or more samples in J, it is added to S. It then searches J for samples in which f is absent. If it finds any, it removes those samples from J and adds f to S; otherwise, it does nothing. The procedure continues until either J or F is empty. If J is empty, then S is a unique metagenomic code for i; otherwise, no unique metagenomic code is constructed for i. Optionally, to improve robustness to noise and reduce false positives, features continue to be added to S after J becomes empty and before F is empty, until a minimum code size is reached. To increase feature diversity, after adding f to S, the remaining features in F with similar presence/absence profiles, defined by a Jaccard similarity cut-off, are removed.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

A query sample is identified as a match to a known sample if all features in the known sample's metagenomic code are re-detected in the query sample using the relaxed detection threshold.

The approach was evaluated using four metagenomic feature types, namely OTUs, Species, Markers, and kbwindows in [FHM+15], to demonstrate its feasibility.

GePMI

GePMI, short for "Generating inter-individual similarity distribution for Personal Microbiome Identification" [WLW+18], matches samples based on the statistical significance of the similarity between them.

For each known sample, pairwise similarity scores with all other samples are computed based on a selected distance measure to generate an inter-individual similarity distribution, which is then fitted into a beta distribution.

When matching a query sample to a known sample, the similarity score between the two samples is computed and tested against the fitted distribution of the known sample. If, under the null hypothesis that the two samples originate from different individuals, the resulting p-value falls below the significance threshold, the null hypothesis is rejected, and the known sample is identified as a match to the query sample.

When matching a query sample to multiple known samples, multiple testing is controlled using the false discovery rate (FDR), and p-values are adjusted to q-values using Benjamini and Yekutieli's method (BY) [BY01]. Known samples with q-values that fall below the significance threshold are identified as matches to the query sample.

The approach was evaluated using long k-mers as metagenomic features to demonstrate its feasibility [WLW+18].

The nearest-neighbour method

The nearest-neighbour method [MHE24] matches samples based on their distance, using feature abundance limits to discretise the abundance values rather than representing features as binary (present or absent), as in the metagenomic-code method, and applying a dynamic threshold to define the acceptance condition.

For each sample, feature abundance limits (see Table 3.2) are used to discretise feature abundance values into six levels: 0, 1, 2, 3, 4, 5. Compared with the metagenomic-code method, this provides a more complex feature representation that captures not only the presence or absence of features but also their level of abundance. The feature abundance limits are informed by the detection thresholds and observations of the metagenomic-code method of Franzosa et al.. The lowest limit is used to ignore unstable low-abundance features, while the higher limits split the complete range of values in a geometrically even way.

Let x denote any feature abundance value. The transformation is defined as

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

$$x \leftarrow \begin{cases} 0 & \text{if } x < t_0, \\ i & \text{if } t_{i-1} \leq x < t_i \text{ for some } 1 \leq i \leq 4, \\ 5 & \text{if } x \geq t_4. \end{cases} \tag{3.1}$$

Table 3.2: Feature abundance limits in the nearest-neighbour method. Reproduced from [MHE24].

Feature level	t_0	t_1	t_2	t_3	t_4
Taxon	0.00005	0.0005	0.005	0.05	0.5
Gene	0.005	0.05	0.5	5	50

For each query sample, the nearest known samples are identified via a nearest-neighbour search, yielding a pair of neighbouring samples. Depending on the selected number of neighbours, more than one neighbouring sample pair may be obtained.

As every sample by definition has one closest sample, to reduce false positives, a dynamic threshold is used to determine whether an obtained neighbouring sample pair is accepted. For a neighbouring sample pair (a, b), where a is the known sample and b is the query sample, $\Delta(a, b)$ represents their distance under a selected distance measure. Each distance between a and any other known sample is denoted by d, and the set of all such distances is denoted by $\mathcal{L}_a$. The sample pair (a, b) is accepted only if $\Delta(a, b)$ is smaller than the minimum distance $d \in \mathcal{L}_a$. In addition, a parameter λ is introduced to control the strictness of the acceptance condition. The dynamic thresholding condition is defined as

$$\Delta(a, b) < \lambda \cdot \min\{d : d \in \mathcal{L}_a\} \tag{3.2}$$

A known sample is identified as a match to a query sample if the two samples form a neighbouring sample pair and satisfy the dynamic thresholding condition.

To improve computational efficiency in nearest-neighbour selection, the known samples are organised into a ball tree[1], a search data structure based on a selected distance measure

The approach was evaluated by Mayer et al. [MHE24] using the same four metagenomic feature types as Franzosa et al. [FHM+15], namely OTUs, Species, Markers, and kbwindows, to demonstrate its feasibility.

[1] A ball tree is a binary tree that partitions the data points into multidimensional balls.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

3.1.2 Evaluation of PMI Approaches

The metagenomic-code and nearest-neighbour methods

Dataset [FHM+15] derived the dataset used to evaluate the metagenomic-code method from the Human Microbiome Project (HMP) [Hum12][2], by selecting a cohort according to specific criteria. This dataset is hereafter referred to as the Franzosa dataset. [MHE24] used the same dataset to evaluate the nearest-neighbour method. It represents a subset of participants from the Human Microbiome Project, comprising 60 adults recruited between 2008 and 2013. Participants were sampled at two visits across five body regions (gut, oral cavity, airway, skin, and vagina) and 18 body sites: gut (1), oral cavity (9), airway (1), skin (4), and vagina (3). Women were sampled at 18 body sites and men at 15 body sites. The interval between visits ranged from 35 to 404 days.

Four metagenomic feature types are derived from the HMP sequencing data in the Franzosa dataset, including taxon-level features (OTUs and Species) and gene-level features (Markers and kbwindows), as shown in Table 3.3.

Table 3.3: Metagenomic feature types in the Franzosa dataset. Adapted from [FHM+15]. Units: RPKM = reads per kilobase per million sample reads.

Feature level	Feature description	Short name	Sequencing basis	Units	Body sites	Paired samples per body site
Taxon	Operational taxonomic units	OTUs	16S rRNA gene	Relative abundance	18	25–105
Taxon	Microbial species	Species	WMS	Relative abundance	6	14–50
Gene	Species-specific marker genes	Markers	WMS	RPKM	6	14–50
Gene	Kilobase windows from microbial reference genomes	kbwindows	WMS	RPKM	6	9–45

Evaluation The metagenomic code and nearest-neighbour methods are applied to feature abundance profiles. [MHE24] evaluated both PMI approaches and provided a detailed description of the evaluation methodology. First-visit samples are treated as known samples, and second-visit samples are treated as query samples.

For each PMI approach and evaluation scenario, the output assigns to each query sample a subset of known samples (identified matches) that the method identified as belonging to the same individual. Depending on the PMI approach and parameter settings, a query sample may have zero, one, or multiple identified matches. These are compared with the ground truth, in which a query sample may have no true match or a single true match.

[2]Further information is available at `https://www.hmpdacc.org/overview/`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Each query sample, along with its identified matches, can thus be classified into one of the following linkage outcomes:

- True positive (TP): a query sample has one or multiple identified matches, and all of them are correct according to the ground truth (successful linkage).
- False positive (FP): a query sample has one or multiple identified matches, but all of them are incorrect according to the ground truth (linkage error).
- TP+FP: A query sample has multiple identified matches, of which some are correct, and the others are incorrect according to the ground truth.
- True negative (TN): a query sample has no identified match, and no true match exists in the ground truth (successful linkage).
- False negative (FN): a query sample has no identified match, but one true match exists in the ground truth (linkage error).
- NA: a known sample for which the metagenomic-code method was not able to construct a unique metagenomic code.

Based on the aggregated counts of the linkage outcomes (TP, FP, TN, and FN) across all query samples, the following metrics are used to provide an overall evaluation:

- Precision: the proportion of identified matches that are correct, computed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3.3}$$

- Recall: the proportion of true matches that are identified, computed as

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3.4}$$

- F1-score: the harmonic mean of precision and recall, reflecting the balance between them, computed as

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{3.5}$$

In addition, McNemar's test, a statistical method for assessing the statistical significance of differences between dependent groups with binary outcomes, is used to determine whether one PMI approach outperforms another. To this end, a contingency matrix is constructed as shown in Table 3.4, which summarises the joint results of the two PMI approaches. Specifically, the matrix counts how often the two PMI approaches agree on

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

the correct match (n_{11}), agree on an incorrect match (n_{00}), or disagree such that one approach identifies the correct match while the other does not (n_{01} and n_{10}).

Table 3.4: Contingency matrix of McNemar's test. PMI approaches: MetaC = the metagenomic-code method; NearN = the nearest-neighbour method.

		NearN	
		correct	incorrect
MetaC	correct	n_{11}	n_{10}
	incorrect	n_{01}	n_{00}

The test statistic is defined as

$$\chi^2 = \frac{(|n_{10} - n_{01}| - 1)^2}{n_{10} + n_{01}}. \tag{3.6}$$

The test statistic is then compared to a χ^2 distribution to obtain the p-value at a significance level of $\alpha = 0.05$. Under the null hypothesis, both PMI approaches perform equally well, with the same probabilities for the discordant cases. Rejection of the null hypothesis indicates a statistically significant difference between the PMI approaches.

Results In [FHM+15], for the metagenomic-code method, unique metagenomic codes are more often constructed (that is, a lower number of NA outcome cases is observed) when using gene-level features (markers and kbwindows) than when using taxon-level features (OTUs and species). Gene-level features also achieve higher TP rates, with 52% for both marker and kbwindow features, compared with 15% and 13% for OTU and species features, respectively. Across all feature types and body sites, stool marker features achieve the highest TP rate of 86%, and the TP rate at a typical body site is about 30%. The authors further suggest that microbiome data cannot be assumed to be completely anonymised, as a non-trivial fraction of samples can be accurately traced back to their original sources.

In [MHE24], compared with the metagenomic-code method, the nearest-neighbour method yields greater improvements on taxon-level features than on gene-level features. For OTU features, the nearest-neighbour method shows at least twice the number of the TP count of the metagenomic-code method across almost all body sites, with statistically significant improvements at 14 of the 18 sites. For species features, it shows higher TP rates across all body sites, with statistically significant improvements at four out of six sites. For marker features, TP counts are higher at four out of six body sites, with statistically significant improvements at two sites. For kbwindow features, neither PMI approach shows a clear advantage.

Regarding precision, recall, and F1-score, Table 3.5 presents the mean values over four metagenomic feature types across six body sites. Although the metagenomic-code method

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

achieves higher precision in five out of six body sites, the nearest-neighbour method yields higher TP counts and thus higher recall. This gain is generally accompanied by only a limited increase in FP counts and a corresponding decrease in precision. Nevertheless, the nearest-neighbour method achieves higher F1-scores in five out of six body sites, indicating a better overall balance between precision and recall.

Table 3.5: Mean precision, recall, and F1-score over metagenomic feature types for the metagenomic-code and nearest-neighbour methods. Reproduced from [MHE24]. PMI approaches: MetaC = the metagenomic-code method; NearN = the nearest-neighbour method.

	Ant. Nares		Buccal Mucosa		Tongue Dorsum	
	MetaC	NearN	MetaC	NearN	MetaC	NearN
Precision	0.760	0.481	0.619	0.554	0.747	0.623
Recall	0.214	0.183	0.225	0.246	0.224	0.370
F1-score	0.332	0.262	0.287	0.338	0.305	0.455
	Suprag. Plaque		Stool		Post. Fornix	
	MetaC	NearN	MetaC	NearN	MetaC	NearN
Precision	0.692	0.655	0.946	0.951	0.779	0.679
Recall	0.290	0.375	0.544	0.840	0.418	0.413
F1-score	0.386	0.473	0.660	0.890	0.440	0.505

As a main finding, the nearest-neighbour method achieves a 30% higher average TP rate for stool samples than the metagenomic-code method. The authors further suggest that microbiome identifiability poses greater privacy threats than previously acknowledged.

GePMI

Dataset [WLW+18] assembled the dataset used to evaluate GePMI from five cohorts, namely HMP, MetaHIT, MRA, FMT, and TTV (see Table 3.6), hereafter referred to as the multi-cohort dataset. It comprises 301 participants and 612 stool samples. Among the participants, 146 provided a single sample, and the others provided 2 to 21 samples. The interval between visits ranged from a few days to approximately two years.

Evaluation GePMI is applied to pairwise similarities computed using long k-mer features, that is, fixed-length DNA subsequences extracted from whole-metagenome shotgun sequencing (WMS) sequencing data, based on Jaccard similarity approximated using MinHash. PMI performance is evaluated using different thresholds for similarity, p-values, and q-values, and measured by auROC and auPRC. Over all possible thresholds, the auROC measures the ability to distinguish between positive and negative cases, and the auPRC measures the trade-off between precision and recall. Both auROC and auPRC take values in $[0, 1]$, with larger values indicating better performance.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 3.6: Summary statistics of the multi-cohort dataset. Visits: the number of visits (number of participants with the corresponding visit count).

Cohort	Short name	Participants	Samples	Visits
Human Microbiome Project	HMP	138	248	1(61), 2(49), 3(25), 4(2), 6(1)
Metagenomics of the Human Intestinal Tract	MetaHIT	119	168	1(83), 2(23), 3(13)
Microbiome Reshaping by Antibiotics	MRA	24	72	3(24)
Impact of faecal Microbiota Transplantation on the intestinal microbiome in metabolic syndrome patients	FMT	13	55	1(2), 3(1), 5(10)
Temporal and Technical Variability of human gut metagenomes	TTV	7	69	4(2), 6(2), 14(2), 21(1)

Results For the multi-cohort dataset, GePMI achieves an auROC of 0.9470 and an auPRC of 0.8712 under the p-value criterion, while under the q-value criterion, it achieves an auROC of 0.8818 and an auPRC of 0.8679.

In addition, [WLW+18] compared GePMI with the metagenomic-code method using stool marker features from the Franzosa dataset, comprising 50 participants and 100 samples. For this subset, GePMI achieves a higher TP rate than the metagenomic-code method, reaching 97% versus 86%. The authors conclude that a personal microbiome can be uniquely identified with high accuracy across several different metagenomic datasets.

3.2 Privacy-preserving techniques

3.2.1 Data anonymisation

Data anonymisation [SMDF+20] (**RQ3**) refers to the process of transforming original microdata, i.e. data that consist of records containing information about individual units [Rei05], into anonymised data. Anonymisation shall thus prevent the subjects' identities from being associated with the released data, restricting both identity disclosure and attribute disclosure while preserving analytical utility as much as possible.

Quasi-Identifiers (QIs) [Sam01] are combinations of non-identifying attributes (such as $gender + age + zipcode$) that may unequivocally identify a subject. To prevent reidentification, QIs can be transformed using approaches such as suppression, generalisation, and microaggregation.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

k-anonymity [Sam01] is a privacy model for releasing while safeguarding the anonymity of the individuals represented in the data. It requires that records be indistinguishable from at least $k-1$ other records with respect to QI values. This ensures that the reidentification probability of any individual is at most $1/k$, thereby preventing unequivocal reidentification when $k > 1$. Microaggregation is a means of achieving k-anonymity. It groups similar records into clusters containing at least k records and replaces the attribute values of each record with the corresponding cluster averages, rather than suppressing or generalising QIs. Microaggregation offers several advantages: (i) the granularity of the input data is not decreased, (ii) continuous numerical attributes are not discretised, and (iii) outliers cause less distortion.

[MKH23] evaluated k-anonymity based on micro-aggregation using ten datasets containing samples of gut microbiomes, and analysed the decreased privacy risk on the anonymised dataset as well as the incurred information loss. They conclude that anonymisation is a suitable addition to the toolkit of privacy-preserving measures for microbiome data and may be particularly useful in specific practical scenarios.

3.2.2 Data synthetisation

Synthetic tabular data generation focuses on learning the statistical distribution of a tabular dataset D_e and producing a synthetic tabular dataset D_s that closely approximates it. Consequently, the overall generation task can be formulated as:

$$D_s \leftarrow p_\theta(D_e),$$

where p_θ represents the generation model and $D_e = \{(x_e^i, y_e^i)\}_{i=1}^{N_l}$ denotes a tabular dataset consisting of N_l samples, where x_e^i indicates all feature values of the i-th row, and y_e^i represents its corresponding label. D_e and D_s follow the same structured format defined by a set of feature names $F = \{f_j\}_{j=1}^{d}$, where f_j represents the j-th feature name, and d denotes the total number of features.

Microbiome data have special characteristics that are challenging to model analytically, including sparsity (many taxa have zero abundance in a sample), overdispersion (the variance of read counts is larger than what is assumed from the usual parametric models), and compositionality (the read counts in a sample sum to a constant) [ZNS+23].

[HME22] conducted an empirical evaluation of two open-source data synthetisation tools, the SDV and synthpop, for machine learning tasks on several publicly available microbiome datasets. Both the SDV and synthpop performed well on the considered tasks, with AUC scores mostly deviating by about ±3% from those on the original data.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

CHAPTER 4

Dataset

This chapter provides a detailed description of the dataset used in this thesis. The Data Understanding section describes the dataset selection criteria, the study and data collection underlying the selected dataset, and the rationale for its selection. The Data Preprocessing section outlines the preprocessing workflow tailored to the characteristics of microbiome data and summarises the statistics of the resulting dataset. Since different evaluations require input data with varying content and structure, the specific data preparation procedures for each modelling task are included in the Experiment Design chapter.

4.1 Data Understanding

This section begins by defining the criteria for selecting a suitable dataset for our analysis, then introduces the selected dataset, and finally explains the rationale for its selection. Although several candidate datasets were considered, only the selected dataset is presented in this section.

4.1.1 Dataset selection criteria

For RQ1, PMI approaches are applied in settings that are more challenging than those considered in previous studies. These PMI settings are defined by the characteristics of the dataset used in this thesis and form the basis for the dataset selection criteria.

To enable the application of PMI, a fundamental requirement is that the dataset contains longitudinal microbiome samples, that is, microbiome samples collected from the same individual at multiple time points. To define more challenging PMI settings, the datasets used in previous studies serve as a baseline. Dataset selection is guided by the following characteristics:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

1. A more homogeneous study population, reflected in demographic factors such as gender, age, geographic location, and ethnicity. Since PMI assumes that intra-individual variation is smaller than inter-individual variation, increased homogeneity in these factors may reduce inter-individual variation and make the PMI setting more challenging.

2. Greater longitudinal diversity, in terms of the number of sampling time points and the time intervals between them. Longer time intervals may increase intra-individual variation and make the PMI setting more challenging. In addition, multiple sampling time points with diverse time intervals and clear temporal metadata enable a more comprehensive comparison across temporal scales.

3. A higher number of true negatives, defined as samples that have no corresponding match in the ground truth due to the absence of sampling at certain time points. Since such cases are common in real-world settings, this allows us to evaluate the ability of PMI approaches to correctly identify unmatched samples. Although true negatives can be simulated by removing samples and are therefore not a strict dataset requirement, their natural occurrence allows participants with only one sampling time point to be included without reducing the total sample size.

4.1.2 MLVS faecal microbiome dataset

The Health Professionals Follow-up Study (HPFS)[1] is an ongoing prospective cohort study that began in 1986 and includes 51,592 U.S. male podiatrists, dentists, osteopathic physicians, veterinarians, pharmacists, and optometrists aged 40–75 years. Participants return questionnaires every 2–4 years with a follow-up rate exceeding 90%, providing information on lifestyle and dietary factors, medication use, and diagnoses of colorectal cancer and other diseases.

Nested within the HPFS cohort, the Men's Lifestyle Validation Study (MLVS) is a sub-study conducted in 2011–2012 to validate self-reported lifestyle and dietary information. The study includes 700 participants aged 52–81 years who provided blood samples in 1994–1995 and completed a food frequency questionnaire (FFQ) in 2010. Participants with a history of coronary artery disease, stroke or transient ischaemic attack, cancer (except squamous or basal cell skin cancer), or major neurological diseases (including amyotrophic lateral sclerosis, Alzheimer's disease, Parkinson's disease, epilepsy, or multiple sclerosis) are excluded from the study. The reasons for these exclusions are not further specified by the creators of the HPFS.

The dataset selected for this thesis, hereafter referred to as the MLVS faecal microbiome dataset, is derived from [MAAD+18]. It represents a subset of participants from the MLVS cohort, comprising 308 individuals recruited between July 2012 and July 2013. The study population consists of healthy, mostly Caucasian health professionals aged 65–81 years (median age: 69 years). Stool samples were collected by participants using a

[1] https://hsph.harvard.edu/research/health-professionals.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

previously validated self-sampling method [LSG+12]. One set of samples was collected 24–72 hours apart, followed by a second set collected approximately six months later. The timelines and sampling time points are shown in Figure 4.1. In total, whole-metagenome shotgun (WMS) sequencing data are available for 929 samples.

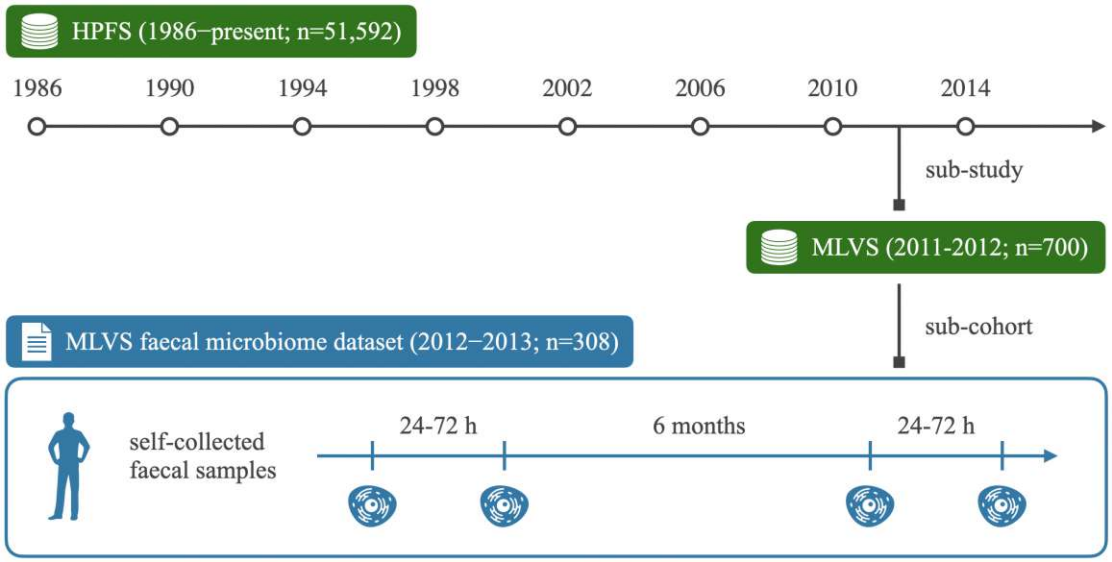


Figure 4.1: Timelines and sampling time points of the HPFS cohort, the MLVS cohort, and the MLVS faecal microbiome dataset. n = number of participants.

4.1.3 Dataset selection rationale

In the studies accompanying the PMI approaches, the Franzosa dataset is used to evaluate the metagenomic-code method [FHM+15] and the nearest-neighbour method [MHE24], while GePMI [WLW+18] is evaluated using a dataset assembled from multiple cohorts (HMP, MetaHIT, MRA, FMT, and TTV). Details of these datasets are provided in Chapter 3.

Table 4.1 compares key characteristics of three datasets: two used in previous studies as baselines (the Franzosa dataset and the multi-cohort dataset) and the one selected for this thesis (the MLVS faecal microbiome dataset). Although the Franzosa dataset includes four metagenomic feature types (described in Chapter 3), the comparison considers only the subset available across all datasets, namely species features extracted from raw WMS data. The table is organised into three groups corresponding to the dataset selection criteria: study population homogeneity, longitudinal diversity, and the number of true negatives.

In terms of study population homogeneity (first row group), the MLVS dataset has the largest number of participants among the three datasets. Although the gender and age of participants are not reported for the multi-cohort dataset, the MLVS dataset has a

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 4.1: Comparison of the Franzosa dataset, the multi-cohort dataset, and the MLVS faecal microbiome dataset with respect to study population homogeneity, longitudinal diversity, and the number of true negatives. "?" = not reported in the original source. Samples by time point: time point (number of samples with the corresponding time point).

	Franzosa [FHM+15]	Multi-cohort [WLW+18]	MLVS (this thesis)
Participants	60	301	308
Sampling body site	multiple	stool	stool
Gender	male and female	?	male
Age	≥18 y	?	65–81 y
Samples	412	612	919
Sampling time points per participant	2	1–6, 14, 21	1–4
Samples by time point	t_1:206, t_2:206	t_1:301, t_2:155, t_3:83, t_4:20, t_5:16, t_6:6, t_{7-14}:3, t_{15-21}:1	t_1:209, t_2:207, t_3:254, t_4:249
Time interval	days–months	days–two years	days–months
Temporal metadata	not available	partial available	available
True negatives (%)	0%	31%	26%

more homogeneous study population than the Franzosa dataset, owing to its narrower age range and all-male composition.

Regarding longitudinal diversity (second row group), the Franzosa dataset represents a classic PMI setting, with paired samples collected from the same participants and body sites at two time points. Although the multi-cohort dataset has the largest number of sampling time points, the samples are distributed with high imbalance, with 301 of 612 samples ($\approx 50\%$) collected at the first time point. Since 146 participants provided only a single sample, at most 155 paired samples are available for evaluating PMI. In contrast, the MLVS dataset has four sampling time points with a relatively balanced sample distribution, and its third and fourth time points have around 25% more samples than the average number of samples by time point in the Franzosa dataset. Moreover, its clear temporal metadata allows separate analyses of short-term variation over days and intermediate-term variation over months. We acknowledge that longer time intervals between sampling time points (> 5 years) would offer additional analytical value, but no suitable dataset with such time intervals was identified among the candidate datasets.

As for the number of true negatives (third-row group), the values represent the proportion of samples with no corresponding match in the ground truth. The Franzosa dataset has no true negatives, whereas the multi-cohort dataset has the highest proportion of true

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

negatives due to the imbalanced distribution of samples across time points. Although the MLVS dataset has a lower proportion than the multi-cohort dataset, the difference remains modest, and, as previously discussed, additional true negatives can be included if necessary by removing samples.

In conclusion, the MLVS dataset has a more homogeneous study population than the other two datasets. Although the multi-cohort dataset has more sampling time points, relatively few individuals have samples from multiple time points. In contrast, the MLVS dataset includes a larger number of individuals with repeated samples and clear temporal metadata, enabling a more meaningful and comprehensive evaluation. Overall, the MLVS dataset meets the dataset selection criteria of this thesis.

4.2 Data Acquisition and Bioinformatics Pipeline

We downloaded the whole-metagenome shotgun (WMS) sequencing data from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA); the dataset is identified as *BioProject accession PRJNA354235*[2].

Table 4.2: Excerpt of *Species* features, for three subjects (columns 24). In each row, the prefixes indicate the taxonomic ranks: `k_` represents kingdom, `p_` phylum, `c_` class, `o_` order, `f_` family, `g_` genus, `s_` species, and `t_` strain. Vertical bars (|) separate the ranks. For example, `k_Bacteria|p_Bacteroidota|c_Bacteroidia|o_Bacteroidales|f_Rikenellaceae|g_Alistipes|s_Alistipes putredinis|t_SGB2318` represents strain SGB2318, which belongs to the species Alistipes putredinis, genus Alistipes, family Rikenellaceae, order Bacteroidales, class Bacteroidia, phylum Bacteroidota, and kingdom Bacteria. The profile contains features from all taxonomic ranks; therefore, some rows represent kingdom-level features, while others represent more specific ranks, such as species or strain.

Clade Name	SRR5155745	SRR5155746	SRR5155747	SRR5155748
k_Bacteria	84.883	94.637	86.853	93.625
k_Eukaryota	0.000	0.230	0.000	0.000
k_Archaea	0.000	0.000	0.445	0.019
k_Bacteria|p_Firmicutes	54.130	59.784	55.274	49.381
k_Bacteria|p_Bacteroidota	24.248	29.818	27.541	41.905
k_Bacteria|p_Pseudomonadota	2.884	0.061	0.293	0.838
k_Bacteria|p_Firmicutes|c_Clostridia	46.742	51.159	49.364	41.299
k_Bacteria|p_Bacteroidota|c_Bacteroidia	22.971	29.337	27.541	41.656
k_Bacteria|p_Pseudomonadota|c_Gammaproteobacteria	2.101	0.061	0.005	0.000
k_Bacteria|p_Firmicutes|c_Clostridia|o_Lachnospirales	23.439	26.088	14.399	24.140
k_Bacteria|p_Bacteroidota|c_Bacteroidia|o_Bacteroidales	22.971	29.337	27.541	41.656
k_Bacteria|p_Firmicutes|c_Clostridia|o_Eubacteriales	22.843	25.070	34.903	17.159
k_Bacteria|p_Firmicutes|c_Clostridia|o_Lachnospirales|f_Lachnospiraceae	23.439	25.401	14.387	23.976
k_Bacteria|p_Firmicutes|c_Clostridia|o_Eubacteriales|f_Oscillospiraceae	17.789	21.031	23.507	14.714
k_Bacteria|p_Bacteroidota|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae	11.545	25.720	14.413	20.735
k_Bacteria|p_Bacteroidota|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides	10.092	21.161	11.444	15.544
k_Bacteria|p_Firmicutes|c_Clostridia|o_Eubacteriales|f_Oscillospiraceae|g_Faecalibacterium	8.244	14.483	6.664	4.236
k_Bacteria|p_Firmicutes|c_Clostridia|o_Lachnospirales|f_Lachnospiraceae|g_Agathobacter	8.161	5.392	0.053	9.645
k_Bacteria|p_Firmicutes|c_Clostridia|o_Lachnospirales|f_Lachnospiraceae|g_Agathobacter|s_Agathobacter rectalis	8.161	5.392	0.053	9.645
k_Bacteria|p_Firmicutes|c_Clostridia|o_Eubacteriales|f_Oscillospiraceae|g_Faecalibacterium|s_Faecalibacterium prausnitzii	7.385	13.881	5.448	0.633
k_Bacteria|p_Bacteroidota|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bacteroides uniformis	6.134	13.078	7.419	2.773
k_Bacteria|p_Firmicutes|c_Clostridia|o_Lachnospirales|f_Lachnospiraceae|g_Agathobacter|s_Agathobacter rectalis|t_SGB4933	8.161	5.392	0.053	9.645
k_Bacteria|p_Bacteroidota|c_Bacteroidia|o_Bacteroidales|f_Bacteroidaceae|g_Bacteroides|s_Bacteroides uniformis|t_SGB1836	6.134	13.078	7.419	2.773
k_Bacteria|p_Bacteroidota|c_Bacteroidia|o_Bacteroidales|f_Rikenellaceae|g_Alistipes|s_Alistipes putredinis|t_SGB2318	3.928	0.000	1.099	0.814

[2]The raw WMS data are available at `https://www.ncbi.nlm.nih.gov/bioproject/PRJNA354235`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The bioinformatics pipeline, including quality control and taxonomic profiling, was set up and executed according to the procedures described in the study accompanying the dataset [MAAD+18]. Quality control was conducted using KneadData[3], involving the removal of human sequences, quality trimming, and depletion of duplicate reads. Taxonomic profiling, a reference-based classification method, was conducted using MetaPhlAn[4] to characterise the taxonomic composition of the samples and generate feature abundance profiles. The resulting feature abundance profile is structured as a table, as shown in Figure 4.2, with samples as columns and mixed-rank features as rows spanning eight taxonomic ranks: Kingdom, Phylum, Class, Order, Family, Genus, Species, and Strain. The values represent the relative feature abundances within each taxonomic rank for each sample.

Since relative feature abundances are meaningful only within the same taxonomic rank and each evaluation uses one rank at a time, the mixed-rank profile was stratified into eight rank-specific profiles. Figure 4.2 illustrates the stratification process from the perspective of feature rows. In each resulting profile, the relative feature abundances within each sample column sum to one. The number of detected features at each taxonomic rank, that is, the number of rows in each rank-specific feature abundance profile, is presented in Table 4.3. Consistent with previous studies, the species-level relative feature abundance profile serves as the primary dataset used in this thesis.

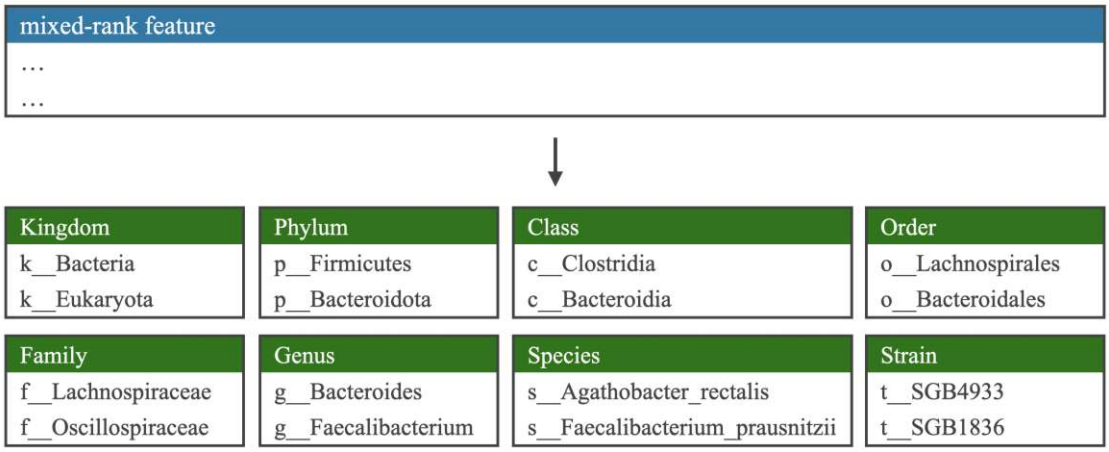


Figure 4.2: Stratification of the mixed-rank profile into rank-specific profiles. The mixed-rank feature abundance profile shown in Figure 4.2 was stratified into eight rank-specific profiles, each containing features from exactly one taxonomic rank, with two example features shown for each rank.

The raw WMS data obtained from the SRA comprise 929 samples, accompanied by metadata describing participant identifiers and sampling time points (first, second, third, and fourth visits). During taxonomic profiling, eight samples failed processing and

[3]Software available at `https://huttenhower.sph.harvard.edu/kneaddata`.
[4]Software available at `https://huttenhower.sph.harvard.edu/metaphlan`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 4.3: Number of detected features by taxonomic rank.

Taxonomic rank	Feature count
Kingdom	4
Phylum	21
Class	416
Order	434
Family	480
Genus	1722
Species	2722
Strain	2821

were excluded. An additional two samples were removed because their sampling-time-point annotations did not correspond to any of the four time points defined in the study accompanying the dataset [MAAD+18]. The preprocessing workflow and the corresponding number of samples, as well as the samples removed at each step, are illustrated in Figure 4.3. The list of removed samples can further be found in Table 1 in the Appendix.

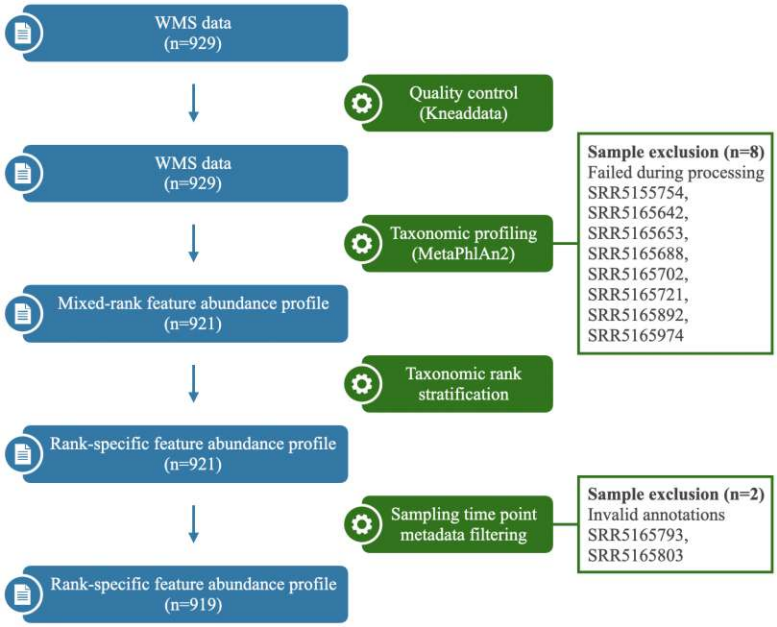


Figure 4.3: Preprocessing workflow and number of samples by step. n = number of samples.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 5

Experiment Design

This chapter provides a detailed description of the experiments conducted in this thesis. Each section (Personal Microbiome Identification, Linkage Errors, and Privacy-Preserving Techniques) is structured around a research question and presents the experimental evaluation that forms the basis for answering it. Within each section, the data preparation outlines the data structure and relevant summary statistics. This is followed by a discussion of the techniques applied, each with an introduction and experimental details, including algorithmic aspects, inputs and outputs, and parameter settings. The evaluation metrics and corresponding methods are then described, while the evaluation results are presented in Chapter 6.

5.1 Personal Microbiome Identification

This section addresses RQ1 by comparing PMI performance across different PMI approaches, longitudinal diversity, and the absence or presence of true negatives. Three PMI approaches, namely the metagenomic-code method (MetaC), GePMI (GE), and the nearest-neighbour method (NearN), are evaluated using the MLVS faecal microbiome dataset. Each PMI approach, as described in Section 3.1, is evaluated in a separate subsection, with ample experimental details. Evaluation scenarios that combine longitudinal diversity with the absence or presence of true negatives are described in the data preparation subsection.

5.1.1 Data preparation

The species-level relative feature abundance profile derived from the MLVS faecal microbiome dataset is used in this experiment, consistent with the original works on these methods [FHM+15, WLW+18, MHE24]. Evaluation scenarios are defined by (i) visit combinations, which determine the assignment of known samples and query samples, as

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

well as longitudinal diversity, and (ii) matching modes, which determine the absence or presence of true negatives.

Regarding visit combinations, known samples and query samples are assigned based on participants' sampling time points. Each participant has up to four samples associated, ordered chronologically by visit. The time intervals between the first and second visits and between the third and fourth visits are 24–72 hours (short-term interval, ST), whereas the time interval between the second and third visits is approximately six months (intermediate-term interval, IT). In total, six visit combinations are considered, each corresponding to a type of time interval (ST or IT), as shown in Table 5.1.

Table 5.1: Visit combinations.

	v1:v2	v1:v3	v1:v4	v2:v3	v2:v4	v3:v4
Known sample	1st visit	1st visit	1st visit	2nd visit	2nd visit	3rd visit
Query sample	2nd visit	3rd visit	4th visit	3rd visit	4th visit	4th visit
Time interval	ST	IT	IT	IT	IT	ST

Regarding matching modes, two settings are considered. In *one-to-one matching*, where each known sample corresponds to exactly one query sample, *paired samples* are constructed based on visit pairs, with equal numbers of known samples and query samples. Only the consecutive visit pairs v1:v2, v2:v3, and v3:v4 are considered, as they provide representative short- and intermediate-term baseline scenarios. Summary statistics for paired samples are provided in Tables 5.2.

Table 5.2: Summary statistics of paired samples by visit pair (one-to-one matching).

Visit pair	Sample pair count
v1:v2	200
v2:v3	156
v3:v4	244

In *matching including true negatives*, where a known sample may not have a corresponding query sample or vice versa, *all samples* collected at each visit are used. All six visit combinations are considered, including the non-consecutive visit pairs v1:v3, v1:v4, and v2:v4, which have higher proportions of true negatives and enable investigation of their effect on PMI performance. Summary statistics for all samples, presented as the number of samples per visit, are provided in Table 5.3, while the proportions of true negatives are reported in Table 5.4.

By combining the visit combinations and matching modes described above, a total of nine scenarios are defined. Relevant summary statistics are presented in Table 5.4. Since PMI approaches match query samples to known samples, the number of true negatives is determined with respect to the query samples, that is, depending on whether each query sample has a known sample as its ground-truth match.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 5.3: Summary statistics of all samples by visit (matching including true negatives).

Visit	Sample count
1st visit	209
2nd visit	207
3rd visit	254
4th visit	249

Table 5.4: Evaluation scenarios and relevant summary statistic. Matching mode: 1:1 = one-to-one matching; +TN = matching including true negatives. Time interval: ST = short-term; IT = intermediate-term.

	1	2	3	4	5	6	7	8	9
Matching mode	1:1	1:1	1:1	+TN	+TN	+TN	+TN	+TN	+TN
Visit combination	v1:v2	v2:v3	v3:v4	v1:v2	v1:v3	v1:v4	v2:v3	v2:v4	v3:v4
Time interval	ST	IT	ST	ST	IT	IT	IT	IT	ST
Known samples	200	156	244	209	209	209	207	207	254
Query samples	200	156	244	207	254	249	254	249	249
True negatives	–	–	–	7	95	91	98	94	5
True negatives (%)	–	–	–	3%	37%	37%	39%	38%	2%

5.1.2 The metagenomic-code method

The algorithm[1] proposed by Franzosa et al. [FHM+15], detailed in section 3.1.1, consists of two phases. In the encoding phase, a metagenomic code, that is, a set of discriminative features, is constructed for each known sample. In the decoding phase, query samples are matched to known samples based on their constructed metagenomic codes.

For encoding, the input is a feature abundance matrix of known samples, with features as rows and samples as columns. The algorithm is applied in relative abundance mode with default parameter settings, including Jaccard similarity cut-off of 0.8, minimum code size of 7, abundance detection and nondetection thresholds of 10^{-3} and 10^{-5}, respectively, and feature ranking based on abundance gap. The output is a set of metagenomic codes for known samples, where a subset of features is assigned to each known sample.

For decoding, the input consists of a set of metagenomic codes for known samples and a feature abundance matrix of query samples, with features as rows and samples as columns. The output is a set of matches for query samples, where a subset of known samples is assigned to each query sample.

The original algorithm considers only the one-to-one matching. In this setting, each iteration processes one known sample and identifies its matches among the query samples.

[1] Python implementation available at `http://huttenhower.sph.harvard.edu/idability`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Any known sample with no identified match is classified as FN. In our setting, however, under the matching scenarios that include true negatives, a known sample may not have a corresponding query sample. In this case, the absence of an identified match is the correct outcome, reflecting the algorithm's ability to identify true negatives. Moreover, since the algorithm iterates over known samples rather than query samples, query samples without corresponding known samples are excluded from evaluation.

Applying the original algorithm in the cases described above would result in incorrect evaluation. To address this issue, we modify the algorithm to iterate over query samples and identify their matches in the known set. During evaluation, if no match is identified for a query sample, the algorithm first determines if the case constitutes TN, that is, if no ground-truth match exists for the query sample, before classifying it as FN.

5.1.3 GePMI

The algorithm[2] proposed by Wang et al. [WLW+18] takes as input a similarity matrix in which rows and columns correspond to the union of known samples and query samples, with entries representing pairwise similarities. The output is a set of matches for the query samples, with a subset of known samples assigned to each query sample.

Wang et al. [WLW+18] compute pairwise Jaccard similarities using long k-mer features extracted from WMS sequencing data. In this thesis, we instead use species-level feature abundance profiles, following the study introducing the selected dataset [MAAD+18], as developing and validating alternative biological feature representations is beyond the scope of this thesis. These profiles correspond to the Species feature type used by Franzosa et al. [FHM+15]. We compute pairwise similarities using a selected distance measure while retaining GePMI's statistical framework for identifying matches. This adaptation allows the three PMI approaches to be evaluated using the same feature representation. However, it differs from the original k-mer-based implementation, and the results therefore describe GePMI's performance on species-level abundance profiles rather than reproducing its original evaluation.

To select the best-fitting distance measure for the adapted GePMI approach, we apply GePMI to similarity matrices computed by different distance metrics and select the one with the best PMI performance. The evaluation considers the evaluation scenarios under *matching including true negatives* (matching mode: all samples), as they provide a sufficiently representative view of the dataset. The evaluated distance measures include Jaccard, Canberra, Bray–Curtis, correlation, Manhattan, Euclidean, and Cosine distances[3].

Let S_{known} and S_{query} denote the sets of known samples and query samples, respectively, and define $S = S_{known} \cup S_{query}$. For any pair of samples $a, b \in S$, their feature abundance vectors are given by $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$, where $i = 1, \ldots, n$ index

[2]Python implementation available at `https://github.com/princello/GePMI`.

[3]We use `scipy` (`https://docs.scipy.org/doc/scipy/reference/spatial.distance.html`).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

features. The corresponding vector means are $\bar{a}$ and $\bar{b}$; $\Delta(a, b)$ denotes the distance between the two samples. Prior to computing Jaccard distance, feature abundance vectors are binarised to values in $\{0, 1\}$, where 0 and 1 indicate the absence and presence of a feature, respectively. The distance between samples a and b is then computed in one of the following ways:

$$\Delta_{\text{Jaccard}}(a, b) = \frac{c_{10} + c_{01}}{c_{11} + c_{10} + c_{01}}, \quad \text{where } c_{pq} = \sum_{\substack{i \\ a_i = p,\ b_i = q}} 1, \quad \text{for } p, q \in \{0, 1\}. \tag{5.1}$$

$$\Delta_{\text{Canberra}}(a, b) = \sum_i \frac{|a_i - b_i|}{|a_i| + |b_i|}. \tag{5.2}$$

$$\Delta_{\text{Bray-Curtis}}(a, b) = \frac{\sum_i |a_i - b_i|}{\sum_i |a_i + b_i|}. \tag{5.3}$$

$$\Delta_{\text{correlation}}(a, b) = 1 - \frac{\sum_i (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i (a_i - \bar{a})^2} \sqrt{\sum_i (b_i - \bar{b})^2}}. \tag{5.4}$$

$$\Delta_{\text{Manhattan}}(a, b) = \sum_i |a_i - b_i|. \tag{5.5}$$

$$\Delta_{\text{Euclidean}}(a, b) = \sqrt{\sum_i (a_i - b_i)^2}. \tag{5.6}$$

$$\Delta_{\text{Cosine}}(a, b) = 1 - \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2} \sqrt{\sum_i b_i^2}}. \tag{5.7}$$

Jaccard and Bray–Curtis distances take values in $[0, 1]$. In our case, the Cosine distance also takes values in $[0, 1]$ because the inputs are non-negative. Canberra, Manhattan, and Euclidean distances take values in $[0, \infty)$, whereas correlation distance takes values in $[0, 2]$. Larger values indicate greater distance between the two samples.

The similarity score between samples a and b is then computed as

$$\text{sim}(a, b) = 1 - \Delta(a, b). \tag{5.8}$$

Accordingly, Jaccard, Bray–Curtis, and Cosine similarity scores take values in $[0, 1]$. Canberra, Manhattan, and Euclidean similarity scores take values in $(-\infty, 1]$, whereas correlation similarity scores take values in $[-1, 1]$. Larger values indicate greater similarity between the two samples.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

One assumption underlying GePMI is that similarity scores follow a Beta distribution bounded within the interval $[0, 1]$. Therefore, for Canberra, correlation, Manhattan, and Euclidean distances, the resulting similarity scores are rescaled[4] to fall within this range. For the set of similarity scores computed between all pairs of samples $a, b \in S$, min(sim) and max(sim) denote the minimum and maximum values of these similarity scores, respectively. The similarity score between samples a and b is normalised as

$$\text{sim}'(a, b) = \frac{\text{sim}(a, b) - \min(\text{sim})}{\max(\text{sim}) - \min(\text{sim})}. \tag{5.9}$$

For each distance metric, the resulting similarity scores are organised into a similarity matrix. The algorithm is applied with default parameter settings, including p-value, q-value, and similarity thresholds of 10^{-3}, 10^{-2}, and 0, respectively.

5.1.4 Nearest-neighbour method

The algorithm[5] proposed in the original study [MHE24] takes as input two feature abundance matrices, one for known samples and one for query samples, with features as rows and samples as columns. The output is a set of matches for query samples, where a subset of known samples is assigned to each query sample.

Several parameter settings can be adjusted, including the number of neighbours, feature abundance limits, the distance measure for nearest-neighbour search, and the distance measure and the strictness for dynamic threshold.

Regarding feature abundance limits, feature abundance values are discretised into six levels, ranging from 0 to 5. Depending on the type of metagenomic features, either taxon-level or gene-level options can be selected. Since species-level abundance extracted from WMS sequencing data corresponds to the taxon-level option, the same setting is adopted in this thesis, using the level thresholds of $(t_0, \ldots, t_4) = (0.00005, 0.0005, 0.005, 0.05, 0.5)$. Feature abundance values below 0.00005 are ignored, as taxon-level features tend to be highly unstable at low abundance and may negatively affect PMI performance [FHM+15]. Other parameter settings remain consistent with the defaults, including one neighbour and the correlation distance (see Equation 5.4) for both the nearest-neighbour search and the strict dynamic threshold.

5.1.5 Evaluation methodology

In the following, we will evaluate the effectiveness of the methods, which refers to the extent to which a PMI approach successfully matches query samples to known samples originating from the same individual under a specific evaluation scenario. Evaluation metrics quantify this extent for each PMI approach and evaluation scenario. Statistical

[4] We use `sklearn` (`https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html`).

[5] Implementation available at `https://github.com/sbaresearch/microbiome-linkage`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

tests then assess the significance of differences in metric values among PMI approaches across all evaluation scenarios to determine whether one outperforms the others.

Evaluation metrics

We follow the method of [MHE24], described in Chapter 3, to count true-positive, false-positive, true-negative, and false-negative cases. There are no *NA* cases, defined as known samples for which the metagenomic-code method cannot construct a unique code, because we iterate over query samples rather than known samples. We also evaluate the overall precision, recall, and F1-score.

Statistical tests

Cochran's Q test [Coc50] and McNemar's test [McN47] are statistical methods for assessing the statistical significance of differences across dependent groups with binary outcomes. McNemar's test is designed for two groups (i.e., paired data), while Cochran's Q test generalises the method to more than two groups. As three PMI approaches are evaluated in this experiment, we apply Cochran's Q test as a global test across all PMI approaches, followed by McNemar's test for pairwise comparisons.

Prior to conducting statistical tests, the linkage outcome (TP, FP, TN, or FN) for each query sample is binarised into a response x, defined as

$$x \leftarrow \begin{cases} 1 & \text{if the linkage outcome is TP or TN,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.10}$$

Cochran's Q test[6] takes three input vectors: block (experimental unit), group (treatment), and response (data value). These vectors can be arranged in a table as shown in Table 5.5, to facilitate understanding of the following steps. Let $i = 1, \ldots, r$ be row indices (block, corresponding to the query sample in our case) and $j = 1, \ldots, c$ be column indices (group, corresponding to the PMI approach in our case). The entry x_{ij} denotes the response (see Equation 5.10) for query sample i under PMI approach j. Row totals are denoted by $u_i = \sum_{j=1}^{c} x_{ij}$ and column totals are denoted by $T_j = \sum_{i=1}^{r} x_{ij}$, where $\bar{u}$ and $\bar{T}$ are the corresponding means, respectively.

The test statistic is defined as

$$Q = \frac{(c-1)\sum_j \left(T_j - \bar{T}\right)^2}{\sum_i u_i \left(1 - \frac{u_i}{c}\right)} = \frac{c(c-1)\sum_j \left(T_j - \bar{T}\right)^2}{c\left(\sum_i u_i\right) - \left(\sum_i u_i^2\right)}. \tag{5.11}$$

The numerator measures the variance between columns, with greater weight placed on larger differences. The denominator captures the variance across columns within

[6] We use the R package `RVAideMemoire` (`https://search.r-project.org/CRAN/refmans/RVAideMemoire/html/cochran.qtest.html`).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 5.5: Input data structure of Cochran's Q test.

	PMI Approach				
Query sample	1	2	...	c	Total (u_i)
1	x_{11}	x_{12}	...	x_{1c}	u_1
2	x_{21}	x_{22}	...	x_{2c}	u_2
⋮	⋮	⋮	⋮	⋮	⋮
r	x_{r1}	x_{r2}	...	x_{rc}	u_r
Total (T_j)	T_1	T_2	...	T_c	

rows implied by the randomisation. Rows with identical outcomes across all columns contribute no variance; thus, a larger denominator reflects greater within-row variation and is more informative for the comparison.

The test statistic is then compared to a χ^2 distribution with $c-1$ degrees of freedom (in our case, 2) to obtain the p-value at a significance level of $\alpha = 0.05$. The critical values at different significance levels can be found in Table 2 in the Appendix. Under the null hypothesis, the probability of a correct outcome is equal across all PMI approaches for each query sample. Rejection of the null hypothesis indicates a statistically significant difference among the PMI approaches.

McNemar's test is explained in detail in chapter 3. To evaluate all pairwise comparisons, McNemar's test is applied to each pair of the three PMI approaches, resulting in three comparisons (MetaC–GePMI, MetaC–NearN, and GePMI–NearN). The resulting p-values are then adjusted using the Holm–Bonferroni method [Hol79], a sequentially rejective multiple test procedure that controls the family-wise error rate.

5.2 Linkage Errors

This analysis addresses RQ2 by identifying patterns related to linkage outcomes, particularly linkage errors, at the sample and feature levels. While the methodology is applicable to all PMI approaches, the evaluation will focus on the approach that achieves the best PMI performance. The input and output of PMI evaluation are analysed to identify underlying patterns at the sample and feature levels, which are described in separate subsections below, each outlining the unit of observation and the pattern of interest. Since these analyses are observational, no evaluation metrics are specified in advance.

5.2.1 Data preparation

The input and output of the PMI evaluation are used in this experiment. For each PMI approach and evaluation scenario, they include (i) the species-level relative feature

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

abundance profile used as input to the PMI approaches; (ii) identified pairs, the matches assigned from known samples to query samples by the PMI approaches; and (iii) linkage outcomes, the classification of each identified pair as TP, FP, TN, or FN during evaluation.

(i) is used only in the feature-level analysis, whereas (ii) and (iii) are used in both the sample-level and feature-level analyses.

5.2.2 Sample-level analysis

For each sample, the analysis examines sample interactions, as reflected by linkage outcomes, and method-related factors of the PMI approach to identify relevant patterns.

Sample interactions

Linkage outcomes reflect sample interactions at different levels, including within-sample pairs and across all samples, as outlined below. In the explanations below, a "true pair" consists of two samples from the same individual, as established by the ground truth, while an "identified pair" consists of two samples that a PMI approach identifies as a match. Links identified by a PMI approach are described as "similar" and links not identified are described as "dissimilar":

- TP: the samples in a true pair are similar, while dissimilar to other samples.
- FP: the samples in a true pair are dissimilar, while the samples in an identified pair are similar. When frequent for a sample, it indicates that the sample is similar to multiple samples.
- TN: a sample is dissimilar to all other samples.
- FN: the samples in a true pair are dissimilar, and also dissimilar to other samples.

Method-related factors

In PMI, linkage errors (FP and FN) may arise from host-related and method-related factors. Host-related factors include genetic, demographic, lifestyle, and environmental influences. Given the relatively homogeneous study population in our dataset, substantial lifestyle changes (e.g., diet, antibiotic or probiotic use, and proton pump inhibitor use) are expected to be the primary drivers of linkage outcomes. However, due to budget constraints, the paid-access metadata on these influences are not available within the scope of this thesis. Therefore, this experiment focuses on method-related factors.

In terms of method-related factors, each PMI approach is characterised by distinct feature abundance thresholds, feature representations, and matching mechanisms, as determined by the parameter settings in PMI evaluation. The metagenomic-code method uses binary feature representations and performs matching based on the discriminative feature sets constructed for each known sample. GePMI conducts statistical significance tests for

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

pairwise similarities using Jaccard distance, a measure based on binary absence/presence information. The nearest-neighbour method uses discretised feature representations, performs nearest-neighbour search based on correlation distance, and applies a dynamic distance threshold derived from all known samples.

The method-related factors associated with linkage outcomes of each PMI approach are outlined below:

For TP:

- Metagenomic-code: sufficiently common discriminative features in the true pair.
- GePMI: Jaccard similarity in the true pair reaches statistical significance.
- Nearest-neighbour: correlation distance in the true pair falls below the dynamic threshold.

For FP:

- Metagenomic-code: insufficiently common discriminative features in the true pair, while misleading common discriminative features in the identified pair. When frequent for a sample, it indicates that the sample shares discriminative features with multiple samples.
- GePMI: Jaccard similarity in the true pair does not reach statistical significance, while misleading Jaccard similarity in the identified pair reaches statistical significance. When frequent for a sample, it indicates that the sample's Jaccard similarity with multiple samples reaches statistical significance.
- Nearest-neighbour: correlation distance in the true pair remains above the dynamic threshold, while misleading correlation distance in the identified pair falls below the dynamic threshold. When frequent for a sample, it indicates that the sample's correlation distance to multiple samples falls below the dynamic threshold.

For TN:

- Metagenomic-code: insufficiently common discriminative features with all other samples.
- GePMI: Jaccard similarity with all other samples does not reach statistical significance.
- Nearest-neighbour: correlation distance to all other samples remains above the dynamic threshold.

For FN:

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

- Metagenomic-code: insufficiently common discriminative features in the true pair.
- GePMI: Jaccard similarity in the true pair does not reach statistical significance.
- Nearest-neighbour: correlation distance in the true pair remains above the dynamic threshold.

Pattern of interest

First, we examine the underlying patterns of sample interactions reflected by the linkage outcomes. For this purpose, each identified pair produced by a selected PMI approach is examined to determine whether one or more of the patterns listed below are observed:

- The origin of FP: When a query sample is classified as FP, does this result from the identified match not being its true match, or from the absence of a true match?
- Overlap between FP and TP: if a query sample Q_A's identified known sample K is not its true match (false positive), and at the same time, K is also identified as a match for its true matching query sample Q_B (true positive), then K is matched to multiple query samples. This indicates that the known sample K is similar to multiple query samples rather than exclusively to its true match Q_B.
- True match linkage status in FN: When a query sample is classified as FN, what is the matching status of its true match? Is it an identified (false positive) match for another query sample?
- Participants across evaluation scenarios: For each individual who provided samples, do the linkage outcomes remain consistent across evaluation scenarios, or do they vary?

Second, we examine the underlying patterns in method-related factors of the PMI approaches. For this purpose, the other PMI approaches are also considered to determine whether one or more of the patterns listed below are observed:

- Linkage outcomes across PMI approaches: When the selected PMI approach yields a linkage outcome, what do the corresponding linkage outcomes look like for the other PMI approaches?
- Linkage patterns across PMI approaches: When a pattern listed above is observed for the selected PMI approach, do the other PMI approaches exhibit the same pattern as well?

Finally, we combine the findings above to identify potential interventions or improvements to the PMI approach that could enhance PMI performance.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.2.3 Feature-level analysis

In this analysis, features serve as the units of observation. The analysis examines (i) feature properties such as abundance, prevalence, and stability, (ii) feature contributions to linkage outcomes based on the PMI approach, and (iii) the associations between these two aspects, to identify relevant patterns.

Feature abundance, prevalence, and stability

To characterise the features in our dataset, we describe them by abundance, prevalence, and stability, following the methodology of the original study on this dataset [MAAD+18].

Feature abundance measures how abundant a feature is across all samples, whereas feature prevalence measures how frequently a feature is present across all samples. For a given feature, let $a = (a_1, \ldots, a_m)$ denote its abundance vector across all samples, where $i = 1, \ldots, m$ indexes samples.

Feature abundance and prevalence are defined as

$$\text{Abundance} = \frac{1}{m} \sum_i a_i. \tag{5.12}$$

$$\text{Prevalence} = \frac{1}{m} \sum_i \mathbf{1}(a_i > 0). \tag{5.13}$$

Feature stability measures how consistently a feature's abundance is maintained over time within the same participants, and is quantified based on Bray–Curtis dissimilarity. For a given feature, let $b = (b_1, \ldots, b_n)$ and $c = (c_1, \ldots, c_n)$ denote its abundance vectors across samples from participants at the first and fourth visits, that is, over the longest sampling interval, where $i = 1, \ldots, n$ indexes participants.

Feature stability is defined as

$$\text{Stability} = \frac{1}{n} \sum_i \left(1 - \frac{|b_i - c_i|}{b_i + c_i}\right). \tag{5.14}$$

Abundance is bounded according to the feature representation used by each PMI approach. Prevalence and stability are bounded within the interval $[0, 1]$, with higher values indicating greater prevalence or stability.

Feature contribution

Feature contribution refers to how a feature contributes to the similarity or dissimilarity of a sample pair. Positive contributions decrease the pairwise distance and thus increase

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

similarity, whereas negative contributions increase the pairwise distance and thus increase dissimilarity. The sample pairs are defined by linkage outcomes. To quantify the contribution, the feature contribution score is derived from the PMI approach.

Evaluated sample pairs For each linkage outcome, the evaluated sample pairs and corresponding rationale are outlined as follows:

- TP evaluated pairs: true pairs, that is, pairs of query samples and their true matches, are used to examine why true matches appear similar, i.e., what caused successful linkage.
- FP evaluated pairs: identified pair, that is, pairs of query samples and their identified matches, are used to examine why identified matches appear similar, i.e. what caused the linkage errors.
- TN evaluated pairs: pairs of query samples and their most similar non-matches are used to examine why the most similar sample is not identified as a match, i.e., what caused the successful non-linkage.
- FN evaluated pairs: pairs of query samples and their true matches are used to examine why true matches appear dissimilar, i.e. what caused these types of linkage errors.

Note that each evaluated sample pair consists of one query sample and one known sample. Therefore, the true match, the identified match, and the most similar non-match mentioned above all refer to the known sample in the pair.

Feature contribution score The measures used to calculate feature contribution scores are derived from specific PMI approaches, each based on its matching mechanism. To more accurately reproduce the original algorithm of each PMI approach, its original feature representation is also used to calculate the feature contribution score. Let S_{known} and S_{query} denote the sets of known samples and query samples, respectively. For an evaluated sample pair with $a \in S_{known}$ and $b \in S_{query}$, their feature abundance vectors are given by $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$, where $i = 1, \ldots, n$ index features. The corresponding vector means are $\bar{a}$ and $\bar{b}$, and s_i represents the contribution score of feature i to the evaluated sample pair (a, b).

For the metagenomic-code method, we define the feature contribution score as

$$s_i = \frac{c_i}{\sum_i c_i},$$

$$\text{where} \quad c_i = \begin{cases} 1, & \text{if included in discriminative features,} \\ 0, & \text{otherwise.} \end{cases} \tag{5.15}$$

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

The score encodes whether feature i is included in the discriminative feature set, i.e., whether it appears in any known sample's metagenomic code. Since the numerator is binary and normalised by the total number of discriminative features, the contribution score is bounded within the interval $[0, 1/\text{number of discriminative features}]$, enabling comparison of feature contributions within and across sample pairs.

For GePMI, we define the feature contribution score as

$$s_i = \frac{c_{11}^{(i)} - c_{10}^{(i)} - c_{01}^{(i)}}{c_{11} + c_{10} + c_{01}} \cdot 100,$$

$$\text{where} \quad c_{pq} = \sum_i c_{pq}^{(i)}, \quad p, q \in \{0, 1\},$$

$$\text{and} \quad c_{pq}^{(i)} = \begin{cases} 1, & a_i = p,\ b_i = q, \\ 0, & \text{otherwise}, \end{cases} \quad p, q \in \{0, 1\}. \tag{5.16}$$

The score follows the structure of the Jaccard distance (see Equation 5.1), where a feature contributes positively when present in both samples, negatively when present in only one sample, and does not contribute when absent in both samples. Since the numerator is restricted to the values 1, 0, or -1 and normalised by the total number of informative features in the sample pair, the contribution score is bounded within the interval $[-1/(c_{11} + c_{10} + c_{01}), 1/(c_{11} + c_{10} + c_{01})]$, enabling comparison of feature contributions within and across sample pairs.

For the nearest-neighbour method, we define the feature contribution score as

$$s_i = \frac{(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i (a_i - \bar{a})^2} \sqrt{\sum_i (b_i - \bar{b})^2}}. \tag{5.17}$$

The score follows the structure of the Pearson correlation coefficient (see Equation 5.4), reflecting the direction and magnitude of each feature's contribution. A feature contributes positively when it deviates from the sample mean in the same direction in both samples, and negatively when it deviates in opposite directions. The magnitude of the deviation determines the strength of the contribution. Consistent with the Pearson correlation coefficient, the contribution score is bounded within the interval $[-1, 1]$.

Pattern of interest

First, we examine global correlations of feature abundance, prevalence, and stability. Second, we examine general patterns of feature contributions to linkage outcomes, such as whether the linkage is driven by similarity, dissimilarity, or both. Finally, we focus on

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

subsets of features associated with specific patterns observed within each linkage outcome to assess whether the key features exhibit similar properties.

5.3 Privacy-Preserving Techniques

This section addresses RQ3 through a comparative evaluation of the effectiveness and statistical fidelity of different privacy-preserving techniques. Two privacy-preserving techniques, namely microaggregation for data anonymisation and synthpop for data synthetisation, are applied to the MLVS faecal microbiome dataset to generate anonymised and synthetic data. The methodology for each privacy-preserving technique, as described in Section 3.2, is presented in a separate subsection. Dataset variant assignments are defined by evaluation methods and described in the evaluation subsection.

5.3.1 Data preparation

The data used in this experiment are characterised by two dimensions: taxonomic rank and dataset variant. Taxonomic rank comprises species- and genus-level data depending on the evaluation method, whereas dataset variant comprises original, anonymised, and synthetic data depending on the privacy-preserving technique.

To evaluate the effectiveness of privacy protection, PMI is used as the evaluation method. Consistent with the original studies [FHM+15, WLW+18, MHE24], the species-level relative feature abundance profile derived from the MLVS faecal microbiome dataset serves as the original dataset. Anonymised and synthetic versions of the data are generated from this dataset by μ-ANT [SMDF+20] and synthpop [NRD16], respectively. The data preparation procedure described in PMI evaluation is then applied to both datasets, ensuring that the relevant summary statistics remain unchanged.

To assess statistical fidelity, enterosignatures (ES) decomposition is used. Consistent with the work of Frioux et al.[FAO+23], the genus-level relative feature abundance profile derived from the MLVS faecal microbiome dataset serves as the original dataset. Anonymised and synthetic versions of the data are also generated. No additional data filtering or partitioning is applied to the three datasets, leaving each with 921 samples and 1,722 features.

5.3.2 k-anonymity via micro-aggregation

For achieving k-anonymity via micro-aggregation, we use the implementation provided by μ-ANT. The algorithm[7] introduced in the original work [SMDF+20] takes as input an original dataset and a configuration; the configuration specifies the dataset description, including attribute names, attribute types, and data types, as well as the protection description, including the anonymisation method assigned to each attribute type.

[7]The Java implementation is available at `https://github.com/CrisesUrv/microaggregation-based_anonymization_tool`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

In our case, the original dataset is represented as a feature abundance matrix, with samples as rows and features as columns. Since the original dataset contains only feature abundance values, attribute names correspond to feature names, all attribute types are set to quasi-identifiers, and all data types are set to continuous numeric values. The anonymisation method used is k-anonymity with $k = 5$. The algorithm's output is an anonymised version of the input feature abundance matrix.

5.3.3 Data Synthetisation via Classification and Regression Trees

For generating synthetic data, we use *synthpop* as the tool. The algorithm[8] introduced in the original work [NRD16] is applied using as input the feature abundance matrix of the original dataset, with samples as rows and features as columns. The synthetisation method used is Classification and Regression Trees (CART), as its non-parametric nature does not require distributional assumptions and is therefore well-suited to highly skewed microbiome data. To mitigate the risk of replicating feature abundance values from the original dataset, which arises from the small number of samples in terminal nodes of regression trees, the minimum terminal node size is set to 5. The algorithm's output is a synthetic version of the input feature abundance matrix.

5.3.4 Evaluation

The objective of privacy-preserving techniques in this thesis is to mitigate privacy concerns associated with PMI while preserving the statistical properties of the original dataset, which are typically competing goals. The PMI approaches and the ES decomposition are applied to the anonymised and synthetic datasets generated by the privacy-preserving techniques to evaluate the effectiveness of their privacy protection and statistical fidelity.

Privacy protection effectiveness

Privacy protection effectiveness refers to the extent to which a privacy-preserving technique successfully mitigates the privacy concerns associated with PMI. In PMI evaluation, PMI approaches are applied to the original dataset, and the resulting PMI effectiveness quantifies the privacy concerns associated with PMI and serves as the baseline.

In this evaluation, PMI approaches are applied to the anonymised and synthetic datasets under the same evaluation scenarios as in the PMI evaluation to obtain their PMI effectiveness. For each evaluation scenario in Table 5.4, the known samples are taken from the original dataset, while the query samples are taken from the evaluated dataset, as shown in Table 5.6.

For each privacy-preserving dataset, we assume the PMI effectiveness decreases; the decrease relative to the original dataset quantifies its privacy-protection effectiveness, that is, how much less effective PMI becomes when such a dataset is used.

[8] We use the R package `synthpop` (`https://cran.r-project.org/web/packages/synthpop/refman/synthpop.html`).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Table 5.6: Dataset variant assignment for PMI.

	i	ii	iii
Evaluated dataset	original	anonymised	synthetic
Research question	RQ 1	RQ 3	RQ 3
Known sample	original	original	original
Query sample	original	anonymised	synthetic

Statistical fidelity

The quality of anonymised and synthetic data can be assessed through statistical fidelity and analytical utility. Statistical fidelity concerns the preservation of statistical properties, whereas analytical utility concerns performance in downstream analyses. As described in section 3.2, [HME22] assesses statistical fidelity through feature correlations, while [MKH23] assesses it through information loss. Both studies evaluate task utility through predictive performance.

In this thesis, the evaluation focuses on statistical fidelity because the variables needed for a downstream prediction task are not publicly available. Given the large number of features in microbiome data, [MKH23] note that comparing feature distributions and correlations becomes less feasible. In our setting, we therefore use enterosignatures (ES) decomposition, which captures ecologically meaningful groups of co-occurring bacteria and their functional potential, to assess whether these patterns are preserved across the different dataset variants.

Statistical fidelity refers to the extent to which a privacy-preserving technique successfully preserves the statistical properties of the original dataset. To evaluate this extent, ES decomposition is applied to the original, anonymised, and synthetic datasets to examine changes in compositional structure across datasets.

In [FAO+23], non-negative matrix factorisation (NMF) is used to decompose the human gut microbiome into microbial signatures, termed enterosignatures (ESs). This method captures intrinsic ecosystem information, such as co-occurring taxa, and represents microbiota composition using continuous latent variables, enabling higher-resolution and more realistic characterisation. Moreover, the ESs derived in the original study correspond to biologically meaningful functional groups.

As illustrated in Figure 5.1, the algorithm[9] takes as input a feature abundance matrix V, with features as rows and samples as columns. NMF decomposes the input feature abundance matrix into two matrices: W, representing feature weights in ESs, and H, representing ES presence in samples.

In ES decomposition, $V \approx WH$, where $\hat{V} = WH$ denotes the reconstruction of V. Here, V_{ij} and $\hat{V}_{ij}$ denote the entries in row i and column j of the input matrix and reconstructed

[9] Python implementation available at `https://enterosignatures.quadram.ac.uk`.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

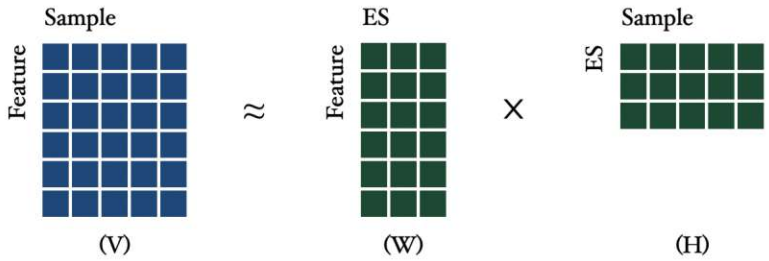


Figure 5.1: Matrix factorisation for ES decomposition.

matrix, respectively. Reconstruction quality is evaluated using three metrics that compare the reconstructed matrix with the input matrix, as defined below. Explained variance measures how much data variation is captured and takes values in $(-\infty, 1]$. Cosine similarity measures similarity in underlying patterns and takes values in $[-1, 1]$, or $[0, 1]$ in our case, due to non-negative data. Reconstruction error quantifies information loss and takes values in $[0, \infty)$. Accordingly, better reconstruction quality is reflected by larger explained variance and cosine similarity, and by smaller reconstruction error.

$$\text{Explained variance} = 1 - \frac{\sum_{i,j}(V_{ij} - \hat{V}_{ij})^2}{\sum_{i,j} V_{ij}^2}. \tag{5.18}$$

$$\text{Cosine similarity} = \frac{\sum_i V_i \hat{V}_i}{\sqrt{\sum_i V_i^2}\sqrt{\sum_i \hat{V}_i^2}}. \tag{5.19}$$

$$\text{Reconstruction error} = \sqrt{\sum_{i,j}(V_{ij} - \hat{V}_{ij})^2}. \tag{5.20}$$

The genus-level relative abundance profile is used because it provides a balanced trade-off between generalisability, functional relevance, and interpretability.

Table 5.7: Dataset variant assignment for ES composition.

	i	ii	iii	iv	v
Decomposition	independent	independent	independent	fixed ES basis	fixed ES basis
Evaluated dataset	original	anonymised	synthetic	anonymised	synthetic
(V)feature×sample	original	anonymised	synthetic	anonymised	synthetic
(W)feature×ES	original	anonymised	synthetic	original	original

The dataset variant assignments are shown in Table 5.7. The combinations examine the ES structure within each dataset and assess how well the original ES structure represents the modified data. In the independent decomposition (columns 1–3), the ES decomposition is performed separately on the original, anonymised, and synthetic

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

datasets (V) to obtain ES compositions (W). In the original study, individual taxa are mostly associated with a single ES, reflecting a distinct pattern of separation. In the decomposition using a fixed ES basis (columns 4–5), the ES composition derived from the original dataset (W) is fixed and reapplied to the anonymised and synthetic datasets (V) to obtain ES distributions across samples (H). Reconstruction quality is evaluated in each scenario, with an explained variance of approximately 0.721 from the dataset provided in the original study, which serves as a baseline.

Table 5.8: Evaluation of the number of ESs.

	3	4	5	6	7	8
Explained variance	0.690	0.727	0.761	0.784	0.798	0.811
Cosine similarity	0.830	0.852	0.872	0.885	0.893	0.900

Regarding parameter settings, we evaluate the number of ESs from 3 to 8 based on the original dataset (see Table 5.8). Based on explained variance and cosine similarity, a value of 5 is selected, as larger values yield only marginal improvements. The NMF is performed using Frobenius divergence, a squared-error loss function, with 50 runs and a maximum of 2000 iterations per run.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 6

Evaluation

This chapter presents the evaluation results for the research questions. Each section Sections 6.1 to 6.3 addresses one of the research questions on Personal Microbiome Identification, Linkage Errors, and Privacy-Preserving Techniques, and follows the corresponding section structure in Chapter 5.

6.1 Personal microbiome Identification

We begin by selecting a distance measure for GePMI based on its PMI performance. Using the selected distance measure, we then compare GePMI with the metagenomic-code and nearest-neighbour methods.

6.1.1 Distance measure selection for GePMI

Table 6.1 reports the PMI effectiveness metrics of GePMI across different distance measures and visit combinations under matching, including true negatives. In terms of F1-score, Jaccard distance achieves the highest values across all distance measures and visit combinations (0.93 in short-term intervals, 0.80 in intermediate-term intervals). It is robust to noise in feature abundance values because it focuses on feature absence/presence, and its bounded range $[0, 1]$ aligns naturally with the beta distribution assumed in GePMI. Among the remaining distance measures, Bray–Curtis and Manhattan distances perform best in short-term intervals (0.83 and 0.82, respectively), whereas Canberra distance performs best in intermediate-term intervals (0.48). Bray–Curtis distance is computed in a similar fashion as Manhattan distance, as it represents a normalised form of the latter, and both are driven by the abundance of dominant features, that is, features with relatively high abundance values. In contrast, the Canberra distance is more sensitive to feature absence/presence and places less emphasis on the abundance of dominant features. Therefore, Jaccard distance is used to compute the input similarity matrix for GePMI in the subsequent comparison.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 6.1: PMI effectiveness by distance measure for GePMI. For each visit combination, the highest F1-score among distance measures is highlighted in green, the second-highest in blue, and the lowest in red.

Visit	Distance	Prec.	Rec.	F1	TP	FP	TP+FP	TN	FN
v1:v2 (ST)	Jaccard	1.00	0.82	0.90	164	0	0	7	36
	Canberra	1.00	0.46	0.63	92	0	0	7	108
	Bray-Curtis	1.00	0.69	0.81	137	0	0	7	63
	correlation	1.00	0.33	0.50	66	0	1	7	133
	Manhattan	1.00	0.65	0.79	130	0	0	7	70
	Euclidean	0.92	0.06	0.11	11	1	2	6	187
	Cosine	1.00	0.30	0.46	60	0	1	7	139
v1:v3 (IT)	Jaccard	0.99	0.65	0.79	104	1	0	94	55
	Canberra	0.95	0.38	0.55	60	3	0	94	97
	Bray-Curtis	0.98	0.25	0.40	40	1	0	94	119
	correlation	0.84	0.10	0.18	16	3	2	94	139
	Manhattan	0.98	0.25	0.40	40	1	0	94	119
	Euclidean	0.46	0.04	0.07	6	7	3	92	146
	Cosine	0.88	0.10	0.17	15	2	2	95	140
v1:v4 (IT)	Jaccard	1.00	0.68	0.81	107	0	0	91	51
	Canberra	0.96	0.32	0.48	50	2	0	90	107
	Bray-Curtis	0.98	0.26	0.41	41	1	0	90	117
	correlation	0.93	0.08	0.15	13	1	1	91	143
	Manhattan	0.98	0.25	0.40	40	1	0	90	118
	Euclidean	0.25	0.01	0.01	1	3	0	89	156
	Cosine	0.93	0.08	0.15	13	1	1	91	143
v2:v3 (IT)	Jaccard	1.00	0.64	0.78	99	0	0	98	57
	Canberra	0.94	0.32	0.47	49	3	0	96	106
	Bray-Curtis	1.00	0.30	0.46	47	0	0	98	109
	correlation	0.85	0.11	0.20	17	3	4	97	133
	Manhattan	1.00	0.30	0.46	47	0	0	98	109
	Euclidean	0.91	0.07	0.12	10	1	0	98	145
	Cosine	0.86	0.12	0.21	18	3	4	97	132
v2:v4 (IT)	Jaccard	1.00	0.67	0.80	104	0	0	94	51
	Canberra	1.00	0.27	0.42	41	0	0	94	114
	Bray-Curtis	1.00	0.26	0.41	40	0	0	94	115
	correlation	0.94	0.11	0.19	16	1	2	94	136
	Manhattan	1.00	0.26	0.41	40	0	0	94	115
	Euclidean	0.67	0.05	0.10	8	4	2	92	143
	Cosine	0.94	0.11	0.19	16	1	2	94	136
v3:v4 (ST)	Jaccard	1.00	0.91	0.95	221	0	1	5	22
	Canberra	0.99	0.59	0.73	141	2	1	5	100
	Bray-Curtis	1.00	0.72	0.84	175	0	1	5	68
	correlation	1.00	0.34	0.51	83	0	3	5	158
	Manhattan	1.00	0.72	0.84	175	0	1	5	68
	Euclidean	0.96	0.22	0.36	53	2	3	5	186
	Cosine	1.00	0.34	0.51	83	0	3	5	158

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

6.1.2 Comparison of PMI approaches

We begin by describing the overall PMI performance under the two matching modes: one-to-one matching and matching including true negatives. The results are then discussed in greater detail in terms of differences among the PMI approaches and longitudinal diversity.

One-to-one matching

The PMI effectiveness metrics are reported in Table 6.3. Precision remains high across all PMI approaches and visit combinations (≥ 0.98). Regarding recall, the nearest-neighbour method achieves the highest values across all PMI approaches and visit combinations (≥ 0.95). Both the metagenomic-code method and GePMI show a notable drop compared to precision, especially in intermediate-term intervals (lowest 0.56). In terms of F1-score, the nearest-neighbour method yields the highest values across all PMI approaches and visit combinations (≥ 0.99 in short-term intervals, ≥ 0.97 in intermediate-term intervals), followed by GePMI (≥ 0.90 in short-term intervals, ≥ 0.80 in intermediate-term intervals) and the metagenomic-code method (≥ 0.94 in short-term intervals, ≥ 0.71 in intermediate-term intervals).

Figure 6.1 shows the proportion of linkage outcomes, where FN (red) consistently exceed FP (orange) across all PMI approaches and visit combinations, indicating that linkage errors are dominated by FN rather than FP; this is also reflected in precision and recall.

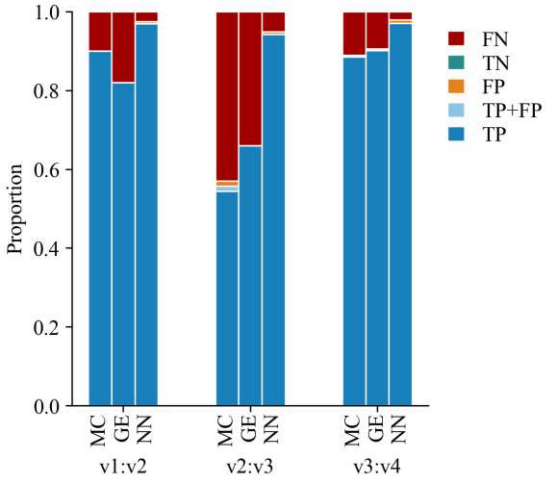


Figure 6.1: PMI linkage outcome proportions by approach on paired samples.

Matching including true negatives

The PMI effectiveness metrics are reported in Table 6.3. Precision remains high across all PMI approaches in short-term intervals (≥ 0.99), but decreases for the metagenomic-code and the nearest-neighbour methods in intermediate-term intervals (lowest: 0.92).

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Regarding recall, the nearest-neighbour method achieves the highest values across all PMI approaches and visit combinations (≥ 0.95). Both the metagenomic-code method and GePMI show a notable drop in precision, especially in intermediate-term intervals (lowest: 0.48). In terms of F1-score, the nearest-neighbour method yields the highest values across all PMI approaches and visit combinations (≥ 0.98 in short-term intervals, ≥ 0.93 in intermediate-term intervals), followed by GePMI (≥ 0.90 in short-term intervals, ≥ 0.78 in intermediate-term intervals) and the metagenomic-code method (≥ 0.94 in short-term intervals, ≥ 0.64 in intermediate-term intervals).

Figure 6.2 shows the proportion of linkage outcomes and includes TN (green), which were absent in Figure 6.1. Comparing FP (orange) and FN (red), more FP are observed for the nearest-neighbour method, while more FN are observed for the metagenomic-code method and GePMI. The nearest-neighbour method tends to identify more matches, whereas the other PMI approaches apply stricter acceptance conditions and are more likely to return no match. As a result, the metagenomic-code method and GePMI produce more TN outcomes, but also more FN outcomes; this is also reflected in precision and recall.

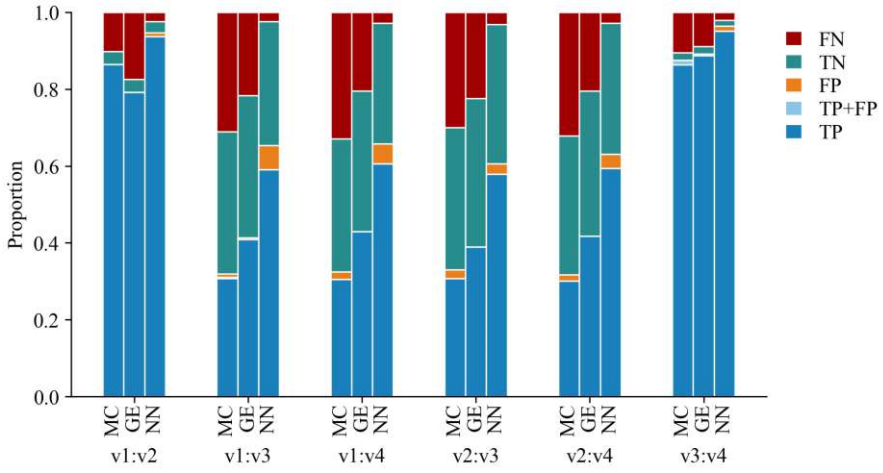


Figure 6.2: PMI linkage outcome proportions by approach on all samples.

Summary

Across the PMI approaches, precision is comparable, while the nearest-neighbour method performs better in recall and F1-score, indicating its ability to identify true matches with fewer incorrect and missed matches. The statistical tests reported in Table 3 in the Appendix support this finding, with the adjusted p-values nearly identical to the original p-values. Cochran's Q test confirms significant differences among the three PMI approaches across all evaluation scenarios. McNemar's test further confirms that the nearest-neighbour method performs best across all evaluation scenarios when pairwise comparing it with the other two PMI approaches. GePMI outperforms the metagenomic-code method with statistically significant differences in four of nine evaluation scenarios,

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

all observed *under matching including true negatives* and intermediate-term intervals, suggesting that this advantage is limited to particular settings.

In the comparison of longitudinal diversity in visit combinations, short-term intervals outperform intermediate-term intervals in precision, recall, and F1-score, as expected. This is due to lower variation in microbiota composition over shorter time intervals, resulting in greater similarity between samples from the same individual taken just a few days apart.

In the comparison between one-to-one matching and matching including true negatives, the change is minimal in short-term intervals and larger in intermediate-term intervals. For the nearest-neighbour method, the average F1-score in intermediate-term intervals decreases from 0.98 under paired samples to 0.94 under all samples, while remaining high despite around 38% true negatives. Overall, the PMI approaches demonstrate feasibility on a dataset more challenging than those used in previous work.

Table 6.2: Average PMI effectiveness by approach. For each evaluation scenario, the highest value among PMI approaches is highlighted in green. Matching mode: 1:1 = one-to-one matching; +TN = matching including true negatives. Time interval: ST = short-term; IT = intermediate-term. PMI Approach: MetaC = the metagenomic-code method; NearN = the nearest-neighbour method.

Mode	Interval	Approach	Prec.	Rec.	F1
1:1	ST	MetaC	1.000	0.895	0.944
		GePMI	1.000	0.863	0.926
		NearN	0.994	0.977	0.985
1:1	IT	MetaC	0.977	0.559	0.711
		GePMI	1.000	0.660	0.795
		NearN	0.993	0.948	0.970
+TN	ST	MetaC	1.000	0.894	0.944
		GePMI	1.000	0.865	0.927
		NearN	0.989	0.977	0.983
+TN	IT	MetaC	0.948	0.492	0.648
		GePMI	0.998	0.659	0.794
		NearN	0.931	0.955	0.943

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 6.3: PMI effectiveness by approach. For each matching mode and visit combination, the highest F1-score among PMI approaches is highlighted in green and the lowest in red.

Mode	Visit	App.	Prec.	Rec.	F1	TP	FP	TP+FP	TN	FN
1:1	v1:v2 (ST)	MetaC	1.00	0.90	0.95	180	0	0	–	20
		GePMI	1.00	0.82	0.90	164	0	0	–	36
		NearN	1.00	0.98	0.99	194	1	–	–	5
1:1	v2:v3 (IT)	MetaC	0.98	0.56	0.71	85	2	2	–	67
		GePMI	1.00	0.66	0.80	103	0	0	–	53
		NearN	0.99	0.95	0.97	147	1	–	–	8
1:1	v3:v4 (ST)	MetaC	1.00	0.89	0.94	216	0	1	–	27
		GePMI	1.00	0.91	0.95	220	0	1	–	23
		NearN	0.99	0.98	0.99	237	2	–	–	5
+TN	v1:v2 (ST)	MetaC	1.00	0.90	0.95	179	0	0	7	21
		GePMI	1.00	0.82	0.90	164	0	0	7	36
		NearN	0.99	0.98	0.98	194	2	–	6	5
+TN	v1:v3 (IT)	MetaC	0.98	0.50	0.66	78	2	1	94	79
		GePMI	0.99	0.65	0.79	104	1	0	94	55
		NearN	0.90	0.96	0.93	150	16	–	82	6
+TN	v1:v4 (IT)	MetaC	0.94	0.48	0.64	76	5	0	86	82
		GePMI	1.00	0.68	0.81	107	0	0	91	51
		NearN	0.92	0.96	0.94	151	13	–	78	7
+TN	v2:v3 (IT)	MetaC	0.93	0.51	0.66	78	6	0	94	76
		GePMI	1.00	0.64	0.78	99	0	0	98	57
		NearN	0.96	0.95	0.95	147	7	–	92	8
+TN	v2:v4 (IT)	MetaC	0.95	0.48	0.64	75	4	0	90	80
		GePMI	1.00	0.67	0.80	104	0	0	94	51
		NearN	0.94	0.96	0.95	148	9	–	85	7
+TN	v3:v4 (ST)	MetaC	1.00	0.89	0.94	215	0	3	5	26
		GePMI	1.00	0.91	0.95	221	0	1	5	22
		NearN	0.99	0.98	0.98	237	3	–	4	5

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

6.2 Linkage Errors

Among all evaluated PMI approaches, the nearest-neighbour method demonstrated the highest PMI effectiveness in PMI evaluation. In this section, we focus on this PMI approach and present the corresponding results.

6.2.1 Sample-level analysis

FP–TP match overlap

As described in section 5.2.2, cases in which a query sample is classified as FP, while its identified match is also classified as TP for another query sample, can occur. This suggests that the identified match is *similar to multiple* query samples, such that, under the nearest-neighbour method, its correlation distance to multiple query samples falls below the dynamic threshold. Table 6.4 presents the summary statistics of FP–TP match overlap in PMI evaluation, specifically the number of query samples classified as FP and, among them, the number whose identified match is also classified as TP for another query sample. For paired samples, the FPs in v1:v2 and v2:v3 overlap completely with TPs. For all samples, four of the six visit combinations exhibit an overlap rate above 50%.

Table 6.4: Summary statistics of TP–FP proposal conflicts. For each evaluation scenario, conflict rates above 50% are highlighted in red. Matching mode: 1:1 = one-to-one matching; +TN = matching including true negatives. Visit combination: time point of known sample to time point of query sample.

Exp. scenario		FP	FP–TP match overlap	
Mode	Visit	Count	Count	Rate (%)
1:1	v1:v2	1	1	100%
1:1	v2:v3	1	1	100%
1:1	v3:v4	2	0	0%
+TN	v1:v2	2	2	100%
+TN	v1:v3	16	9	56%
+TN	v1:v4	13	8	62%
+TN	v2:v3	7	5	71%
+TN	v2:v4	9	4	44%
+TN	v3:v4	3	1	33%
Total		54	31	57%

Improvements to the nearest-neighbour method

To identify potential improvements to the nearest-neighbour method, we consider an extension of the case described above. In this case, a known sample K_A is identified as the nearest neighbour of both query samples Q_A and Q_B, whereas it is the true match

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

only for Q_A. This raises the question of whether K_A is closer to Q_A than to Q_B. If this is the case (Assumption 1), assigning K_A only to its nearest query sample decreases the FP count. This further raises the question of whether the *second nearest neighbour of Q_B is its true match*. If this is also the case (Assumption 2), rematching Q_B to a different known sample further increases the TP count.

Based on these assumptions, a potential improvement using mutual nearest-neighbour (MNN) matching [HLMM18] is explored in the following. A pair of samples from different sampling time points is defined as mutual nearest neighbours if they are each other's nearest neighbours. Two approaches to improvement are proposed, depending on whether the query sample originally classified as FP (Q_B) is rematched. In MNN without rematching (corresponding to Assumption 1), a known sample is assigned to a query sample only if they are mutual nearest neighbours. In MNN with rematching (corresponding to Assumption 2), if a query sample and its nearest known sample are not mutual nearest neighbours, the subsequent nearest known samples are examined iteratively.

The original nearest-neighbour method for PMI is given in Algorithm 6.1 (a more detailed description is provided in Section 3.1. During the nearest-neighbour search (lines 4–5), the nearest neighbour of the query sample is identified, and during the dynamic thresholding (lines 6–10), this nearest neighbour is then examined to determine whether it is accepted as a match for the query sample.

Algorithm 6.1: Nearest-neighbour PMI

Input: Feature abundance matrices for known samples $\mathbf{A} \in \mathbb{R}^{m \times p}$ and query samples $\mathbf{B} \in \mathbb{R}^{n \times p}$, feature abundance transformation $\mathcal{T} : \mathbb{R} \to \mathbb{R}$, distance measure Δ, and threshold scaling parameter λ

Output: Vector of identified matches for query samples $\vec{r}$

1 $\mathbf{A} \leftarrow \mathcal{T}(\mathbf{A})$; // feature abundance limits
2 $\mathbf{B} \leftarrow \mathcal{T}(\mathbf{B})$;
3 **for** $i \leftarrow 1$ **to** n **do**
4 $\quad \mathcal{S}_A \leftarrow \operatorname{argsort}_{1 \le k \le m} \Delta(\mathbf{a}_k, \mathbf{b}_i)$; // indices of A sorted by dist to b_i
5 $\quad c \leftarrow \mathcal{S}_A[1]$; // index of b_i's NN in A
6 $\quad \ell \leftarrow \Delta(\mathbf{a}_c, \mathbf{b}_i)$; // dist between b_i and its NN a_c
7 $\quad \mathcal{L} \leftarrow \{\Delta(\mathbf{a}_c, \mathbf{a}_k) \mid 1 \le k \le m,\ k \ne c\}$; // dist from a_c to other A samples
8 $\quad$ **if** $\ell < \lambda \cdot \min(\mathcal{L})$ **then** // distance comparison
9 $\quad\quad r_i \leftarrow c$;
10 $\quad$ **end**
11 **end**
12 **return** $\vec{r}$;

MNN without rematching The proposed approach is shown in Algorithm 6.2. MNN matching is incorporated into the dynamic thresholding (lines 6–12). When a known

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

sample is identified as the nearest neighbour of a query sample, the threshold is derived from both (1) the pairwise distances between the nearest neighbour and the other known samples, as proposed in the original study, and (2) the pairwise distances between the nearest neighbour and the other query samples, as proposed in this thesis. The latter reflects which query sample is nearest to the known sample. In this way, if a known sample is identified as the nearest neighbour of multiple query samples, it is assigned only to its nearest query sample. Since the other query samples remain unmatched, they will be classified as TN or FN depending on whether they have a true match.

Algorithm 6.2: Nearest-neighbour PMI: MNN without rematching

Input: Feature abundance matrices for known samples $\mathbf{A} \in \mathbb{R}^{m \times p}$ and query samples $\mathbf{B} \in \mathbb{R}^{n \times p}$, feature abundance transformation $\mathcal{T} : \mathbb{R} \to \mathbb{R}$, distance measure Δ, and threshold scaling parameter λ

Output: Vector of identified matches for query samples $\vec{r}$

1 $\mathbf{A} \leftarrow \mathcal{T}(\mathbf{A})$;
2 $\mathbf{B} \leftarrow \mathcal{T}(\mathbf{B})$;
3 **for** $i \leftarrow 1$ **to** n **do**
4 $\quad \mathcal{S}_A \leftarrow \text{argsort}_{1 \le k \le m}\, \Delta(\mathbf{a}_k, \mathbf{b}_i)$;
5 $\quad c \leftarrow \mathcal{S}_A[1]$;
6 $\quad \ell \leftarrow \Delta(\mathbf{a}_c, \mathbf{b}_i)$; // dist between b_i and its NN a_c
7 $\quad \mathcal{L}_A \leftarrow \{\Delta(\mathbf{a}_c, \mathbf{a}_k) \mid 1 \le k \le m,\ k \ne c\}$; // dist from a_c to other A samples
8 $\quad \mathcal{L}_B \leftarrow \{\Delta(\mathbf{a}_c, \mathbf{b}_j) \mid 1 \le l \le n,\ l \ne i\}$; // dist from a_c to other B samples
9 $\quad \mathcal{L} \leftarrow \mathcal{L}_A \cup \mathcal{L}_B$;
10 $\quad$ **if** $\ell < \lambda \cdot \min(\mathcal{L})$ **then** // distance comparison
11 $\quad\quad r_i \leftarrow c$;
12 $\quad$ **end**
13 **end**
14 **return** $\vec{r}$;

MNN with rematching The proposed approach is defined in Algorithm 6.3. A variation of MNN matching is incorporated into the nearest-neighbour search (lines 4–13). The known samples are sorted by their distance to the query sample to identify its nearest known sample. The procedure then iterates over the sorted known samples and applies dynamic thresholding only if the query sample is also the nearest query sample to the known sample under consideration, i.e., they are mutually closest. If the nearest known sample does not satisfy this criterion, the procedure examines the next nearest known sample according to the same criterion, continuing until all known samples have been examined. In this way, if a known sample is the nearest known sample of multiple query samples, the procedure continues only with its nearest query sample, and the other query samples retain the opportunity to be matched to their subsequent nearest known samples.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Algorithm 6.3: Nearest-neighbour PMI: MNN with Rematching

Input: Feature abundance matrices for known samples $\mathbf{A} \in \mathbb{R}^{m \times p}$ and query samples $\mathbf{B} \in \mathbb{R}^{n \times p}$, feature abundance transformation $\mathcal{T} : \mathbb{R} \to \mathbb{R}$, distance measure Δ, and threshold scaling parameter λ

Output: Vector of identified matches for query samples $\vec{r}$

$\mathbf{A} \leftarrow \mathcal{T}(\mathbf{A})$;
$\mathbf{B} \leftarrow \mathcal{T}(\mathbf{B})$;
for $i \leftarrow 1$ **to** n **do**
  $\mathcal{S}_A \leftarrow \text{argsort}_{1 \le k \le m} \Delta(\mathbf{a}_k, \mathbf{b}_i)$; // indices of A sorted by dist to b_i
  for $j \leftarrow 1$ **to** m **do**
    $c \leftarrow \mathcal{S}_A[j]$; // index of b_i's j-th NN in A
    $\mathcal{S}_B \leftarrow \text{argsort}_{1 \le l \le n} \Delta(\mathbf{a}_c, \mathbf{b}_l)$; // indices of B sorted by dist to a_c
    $d \leftarrow \mathcal{S}_B[1]$; // index of a_c's NN in B
    if $i = d$ **then** // b_i is a_c's NN b_d
      $found \leftarrow$ true;
      break;
    end
  end
  if $found$ **then**
    $\ell \leftarrow \Delta(\mathbf{a}_c, \mathbf{b}_i)$;
    $\mathcal{L} \leftarrow \{\Delta(\mathbf{a}_c, \mathbf{a}_k) \mid 1 \le k \le m,\ k \ne c\}$;
    if $\ell < \lambda \cdot \min(\mathcal{L})$ **then**
      $r_i \leftarrow c$;
    end
  end
end
return $\vec{r}$;

Evaluation The original nearest-neighbour method, as well as the improvement approaches including MNN without rematching and MNN with rematching, are applied to the evaluation scenarios from PMI evaluation and evaluated in terms of PMI effectiveness.

The PMI effectiveness metrics are reported in Table 6.5. Across all evaluation scenarios, the summed FP counts for the original method, MNN without rematching, and MNN with rematching are 54, 10, and 48, respectively. In addition, the FP–TP match overlap rate is reduced to zero for both improvement approaches.

The results for MNN without rematching show a large drop in the FP count, which supports Assumption 1: when a known sample is identified as the nearest neighbour of multiple query samples, it remains the nearest neighbour of its true query sample. Moreover, for query samples left without an identified match, this results in a much larger increase in the TN count than in the FN count, indicating that most query samples in the FP–TP match overlap do not have a true match. The results for MNN with rematching

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

show a slight drop in FP count and a slight increase in TP count, which does not support Assumption 2: when a query sample's nearest known sample is the true match of another query sample, its subsequent nearest known sample is not necessarily the true match. This implies that FPs in FP–TP match overlap are driven more by insufficient similarity in the true pair than by the general similarity of the known sample.

In conclusion, MNN without rematching performs well, achieving the highest F1 score in six of nine evaluation scenarios. However, incorporating MNN matching requires multiple query samples and is therefore more suitable for matching between two data sources (the setting considered in this thesis) than when solely matching a single query sample to a data source.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 6.5: PMI effectiveness by improvement strategy for the nearest-neighbour method. For each evaluation scenario, the highest TP and TN counts among improvement strategies are highlighted in blue. Matching mode: 1:1 = one-to-one matching; +TN = matching including true negatives. Visit combination: time point of known sample to time point of query sample. Time interval: ST = short-term; IT = intermediate-term. Improvement strategy: NearN = the original nearest-neighbour method; MNN = Mutual nearest neighbour; MNN-RM = Mutual nearest neighbour with rematching.

Mode	Visit	Strategy	Prec.	Rec.	F1	TP	FP	TN	FN
1:1	v1:v2 (ST)	NearN	1.00	0.98	0.99	194	1	–	5
		MNN	1.00	0.97	0.99	194	0	–	6
		MNN-RM	1.00	0.98	0.99	196	0	–	4
1:1	v2:v3 (IT)	NearN	0.99	0.95	0.97	147	1	–	8
		MNN	1.00	0.94	0.97	147	0	–	9
		MNN-RM	1.00	0.95	0.97	148	0	–	8
1:1	v3:v4 (IT)	NearN	0.99	0.98	0.99	237	2	–	5
		MNN	0.99	0.98	0.99	237	2	–	5
		MNN-RM	0.99	0.98	0.99	237	2	–	5
+TN	v1:v2 (ST)	NearN	0.99	0.98	0.98	194	2	6	5
		MNN	1.00	0.97	0.99	194	0	7	6
		MNN-RM	1.00	0.98	0.99	195	1	7	4
+TN	v1:v3 (IT)	NearN	0.90	0.96	0.93	150	16	82	6
		MNN	0.98	0.95	0.97	150	3	93	8
		MNN-RM	0.93	0.96	0.95	152	11	85	6
+TN	v1:v4 (IT)	NearN	0.92	0.96	0.94	151	13	78	7
		MNN	0.99	0.96	0.97	151	1	90	7
		MNN-RM	0.95	0.96	0.96	152	8	83	6
+TN	v2:v3 (IT)	NearN	0.96	0.95	0.95	147	7	92	8
		MNN	0.99	0.94	0.97	147	1	97	9
		MNN-RM	0.93	0.95	0.94	148	12	86	8
+TN	v2:v4 (IT)	NearN	0.94	0.96	0.95	148	9	85	7
		MNN	0.99	0.96	0.97	148	1	93	7
		MNN-RM	0.93	0.96	0.95	149	11	83	6
+TN	v3:v4 (IT)	NearN	0.99	0.98	0.98	237	3	4	5
		MNN	0.99	0.98	0.99	237	2	5	5
		MNN-RM	0.99	0.98	0.99	237	3	5	4

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

6.2.2 Feature-level analysis

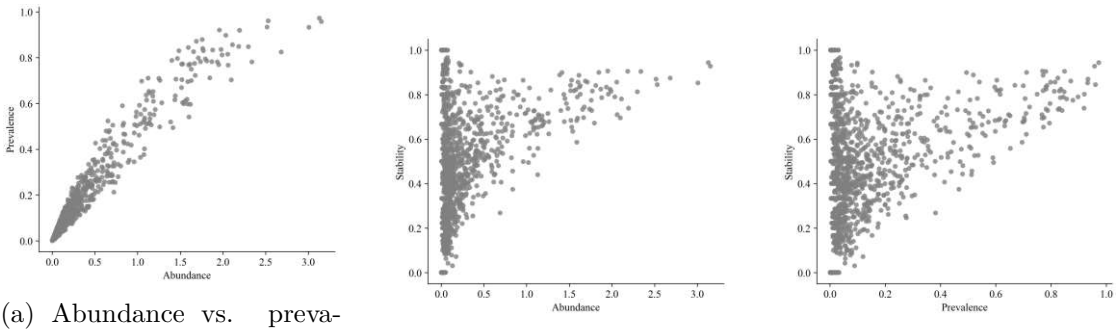
Feature abundance, prevalence, and stability

Based on the summary statistics presented in Table 6.6, feature abundance is highly skewed, with the median substantially lower than the mean. This indicates that only a small number of features are highly abundant, while the majority remain at low abundance. A similar pattern is observed for feature prevalence, where most features appear in a limited number of samples. In contrast, feature stability shows a more moderate distribution, suggesting that a relatively larger proportion of features maintain consistent abundance across sampling time points.

The relationships among these feature characteristics are further illustrated by the cross comparisons in Figure 6.3. Feature abundance and prevalence are positively correlated: highly abundant features tend to be present more frequently across samples. With respect to feature stability, no clear pattern is observed among features with low abundance and prevalence, whereas features with higher abundance and prevalence are generally more stable over time, although the relationship is not strictly linear. Furthermore, the association between feature prevalence and stability appears stronger than that between feature abundance and stability.

Table 6.6: Summary statistics of feature abundance, prevalence, and stability.

	Mean	Median	Min	Max
Feature abundance	0.125	0.017	0.001	3.153
Feature prevalence	0.065	0.011	0.001	0.973
Feature stability	0.422	0.418	0.000	1.000



(a) Abundance vs. prevalence.

(b) Abundance vs. stability.

(c) Prevalence vs. stability.

Figure 6.3: Feature abundance, prevalence, and stability

Feature contribution

Using the feature contribution score derived from the nearest-neighbour method, we evaluate the sample pairs corresponding to each linkage outcome. Extending the definition

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

in section 5.2.3, for true negatives (TNs), the evaluated pairs consist of each query sample and its most similar non-match. This most similar non-match, referred to as the nearest non-match, is defined as the known sample with the highest pairwise similarity to the query sample among all known samples.

Overall feature contribution For each feature, contribution scores are computed across all evaluated sample pairs within each linkage outcome. The resulting scores are then grouped by sign into positive and negative contributions, and the median is derived for each group. Accordingly, each feature is represented within each linkage outcome by two values: the median positive contribution score ($s > 0$), contributing to similarity, and the median negative contribution score ($s < 0$), contributing to dissimilarity.

Table 6.7 reports the summary statistics of feature contribution scores aggregated across all features within each linkage outcome. The large differences between means and medians, together with the high maximum values, further suggest that feature contribution scores are strongly right-skewed, with a small number of features contributing disproportionately. Across all linkage outcomes, positive contribution scores show a much wider spread (0.006 to 23.802) than negative contribution scores (0.023 to 0.825). However, median negative contribution scores consistently exceed median positive contribution scores, despite the lower non-zero counts of the latter. This suggests that while *similarity is driven by a small number of features with large positive contributions*, *dissimilarity is driven by a larger number of features with smaller negative contributions.*

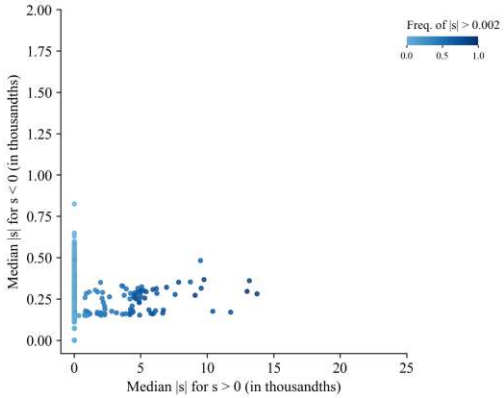
Table 6.7: Summary statistics of overall feature contribution by linkage outcome. The values are reported in thousandths.

	TP		FP		TN		FN	
	$s > 0$	$s < 0$	$s > 0$	$s < 0$	$s > 0$	$s < 0$	$s > 0$	$s < 0$
Mean	0.178	0.204	0.212	0.263	0.178	0.209	0.186	0.183
Median	0.022	0.159	0.024	0.231	0.024	0.160	0.015	0.142
Min	0.021	0.071	0.021	0.121	0.023	0.071	0.006	0.023
Max	13.753	0.825	13.286	0.817	12.915	0.784	23.820	1.809
$\lvert s\rvert > 0$ Count	2637	2134	2637	1346	2637	1463	2637	1025

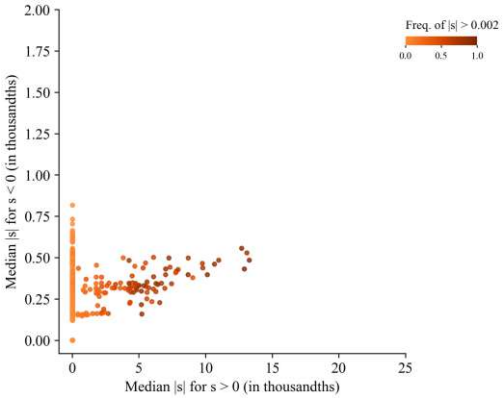
The median contribution scores of each feature are shown in scatter plots, in Figures 6.4a to 6.4d, with positive values on the x-axis and negative values on the y-axis. The figures illustrate the typical magnitudes of positive and negative contributions for each feature across evaluated sample pairs. Colour intensity represents the frequency with which a feature reaches $|s| > 0.002$, with greater intensity indicating more frequent substantial contributions across evaluated sample pairs.

Across all four figures, a thin vertical "band" at $x \approx 0$ indicates that many features exhibit zero positive contribution, while their negative contributions span a range of

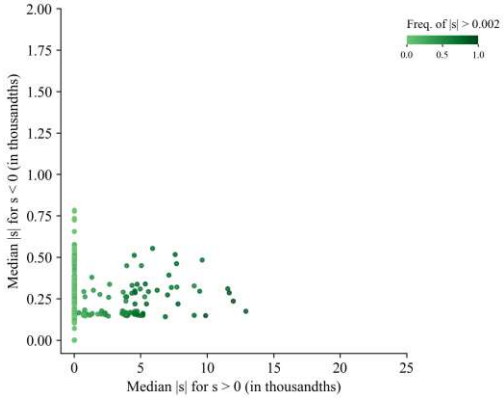
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



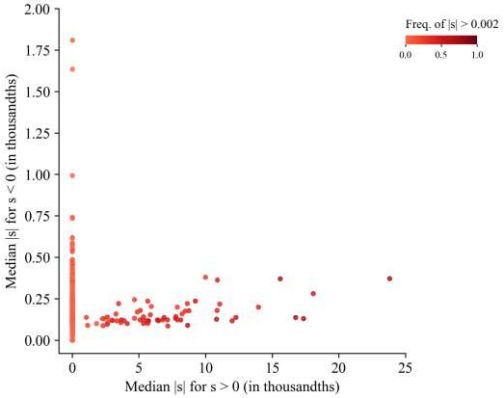
(a) TP evaluated pairs (query sample and correct match pairs).



(b) FP evaluated pairs (query sample and incorrect match pairs).



(c) TN evaluated pairs (query sample and nearest sample pairs).



(d) FN evaluated pairs (query sample and correct match pairs).

Figure 6.4: Overall feature contribution patterns

values. This may explain why median negative scores generally exceed median positive scores in Table 6.7. The small cluster near $x = 0$ and $y = 0$ suggests that features with zero negative contribution show negligible positive contribution, consistent with the small non-zero count of negative scores and the negligible median positive scores reported in the table. Based on the colour intensity, features with more frequent substantial contributions tend to appear mainly when $x > 0.05$, implying that such repeatedly influential features are more likely to occur beyond a certain contribution threshold.

In the TP plot in Figure 6.4a, a "cloud" is visible along the positive contribution axis, extending approximately from $x = 1$ to $x = 15$, while the negative contributions remain within a relatively narrower range. This may serve as a baseline pattern for query samples and their true matches. Compared with the TP plot, the FP plot in Figure 6.4b shows a similar overall pattern. However, the "cloud" is denser and shifted upward, and the thin vertical "band" is also located slightly higher – this suggests larger magnitudes of both

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

positive and negative contributions, with negative contributions being more pronounced, consistent with Table 6.7, where the mean positive score increases from 0.178 to 0.212 in TP, and the mean negative score increases from 0.212 to 0.263 in FP. Therefore, although all of these pairs are identified by the PMI approach, FP pairs appear more dissimilar than TP pairs.

Regarding the TN plot in Figure 6.4c, the "cloud" is spread more vertically upward, especially over the range from $x = 5$ to $x = 10$, indicating a greater negative contribution, while positive contributions remain relatively unchanged. Thus, although dissimilarity increases in both FP and TN pairs relative to TP pairs, only FP pairs are accompanied by an increase in similarity. This may explain why query samples and their nearest samples are filtered out by the dynamic distance threshold in the nearest-neighbour method. In terms of the FN plot in Figure 6.4d, it shows a more "band"-like pattern rather than a "cloud", characterised by higher positive contributions and lower negative contributions. This is consistent with Table 6.7, where the mean positive score increases from 0.178 in TP to 0.186 in FN, while the mean negative score decreases from 0.204 in TP to 0.183 in FN. However, the median positive score decreases from 0.022 in TP to 0.015 in FN, indicating that the increase in the mean is driven by a small number of high positive scores.

Individual species contribution Based on the findings above, frequently influential features were further examined, as reflected by darker colour intensity in the scatter plots. Specifically, features with $|s| > 0.002$ and frequency > 0.5 were selected; that is, these features show an absolute contribution magnitude greater than 0.002 in more than 50% of the evaluated sample pairs. Figure 6.5 presents the species identified as frequently influential features. The first blue block represents their median positive contribution score, the second orange block represents their median negative contribution score, and the three grey bars represent their abundance, prevalence, and stability, respectively.

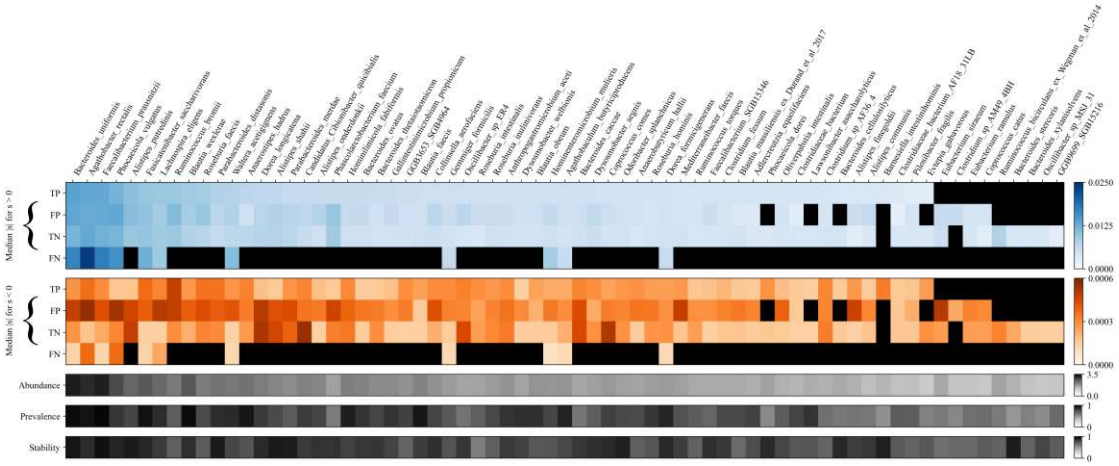


Figure 6.5: Frequently influential features by linkage outcome.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Each linkage outcome exhibits a distinct set of frequently influential features among its evaluated sample pairs, with high overlap across linkage outcomes. In total, with the selection criteria described above, 60 species were identified among the TP evaluated pairs, 59 among the FP evaluated pairs, 67 among the TN evaluated pairs, and 11 among the FN evaluated pairs. Combined across all linkage outcomes, these yielded 69 unique species in the figure.

In the comparison between positive (first, blue block) and negative (second, orange block) contributions, the positive contributions show greater consistency, whereas the negative contributions show greater variation across species and linkage outcomes.

The first four species on the left are notable for high positive contributions, along with high abundance, prevalence, and stability: *Bacteroides uniformis*, *Agathobacter rectalis* (also known as Eubacterium rectale [RKS+16]), *Faecalibacterium prausnitzii*, and *Phocaeicola vulgatus (also known as Bacteroides vulgatus* [CVL+24]). They are all common members of the human gut microbiome, frequently detected in healthy adults [HRW+21, MMR+13].

In the comparison of linkage outcomes, the patterns are consistent with the scatter plots, with TP as the baseline. Compared with TP, FP show increases in both positive and negative feature contributions (darker shading in the blue and orange blocks). TN maintain positive contributions at a level similar to TP's, but exhibit increased negative contributions (similar blue shading but darker orange shading). FN is generally comparable to TP (similar shades of blue and orange); however, certain features exhibit substantially larger contributions than those observed in any other linkage outcome (only 11 features are identified as frequently influential features).

The frequently influential features in FP-evaluated pairs include four species that are not identified in TP-evaluated pairs: *Eubacterium siraeum*, *Clostridium sp. AM49 4BH*, *Eubacterium ramulus*, and *Coprococcus catus*. These species also show lower abundance, prevalence, and stability. Among them, *E. siraeum*, *E. ramulus*, and *C. catus* are associated with gut fermentation-related processes, including wheat bran fermentation and xylan degradation in *E. siraeum*, flavonoid degradation in *E. ramulus*, and lactate utilisation in *C. catus* [DRQ+16, SPJB02, SLT+22]. This may reflect greater variation in these taxa across individuals and over time, possibly driven by dietary factors, thereby increasing their likelihood of contributing to FP-evaluated pairs and generating mixed signals. Notably, three of these taxa are also observed in TN evaluated pairs.

For FN, as mentioned, Figure 6.4d provides some indication that the true matches could potentially have been identified, based on the higher positive contributions and lower negative contributions. However, Figure 6.5 shows that only a small number of features are frequently influential. Overall, too few features contribute sufficiently strongly to support the linkage.

There are still several patterns that can be identified, although the underlying reasons remain unclear. For example, *Phascolarctobacterium faecium* shows a strong similarity contribution in both FP and TN pairs, both of which involve non-true matches. In

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

addition, *Alistipes putredinis*, one of the most abundant taxa, is not observed in FN pairs, even though it is usually present in true-match pairs.

6.3 Privacy-Preserving Techniques

Privacy protection effectiveness

The PMI effectiveness metrics for the original dataset are reported in Table 6.3, and those for the anonymised and synthetic datasets are reported in Tables 4 and 5 in the Appendix. Across all PMI approaches and evaluation scenarios, a total of 2,062 mapping attempts are performed, of which 18 TPs are observed in the anonymised dataset and 1 TP for the synthetic dataset, compared with 1,605 TPs for the original dataset, reflecting a substantial decrease in successful linkages. Notably, the 19 mapping attempts classified as TP in the anonymised and synthetic datasets are also classified as TP in the original dataset. Given the negligible TP count, TN linkage outcomes are not discussed further.

Among the 18 TPs observed in the anonymised dataset, the top 10 taxa ranked by median absolute contribution score are further examined. Table 6.8 presents their median positive contribution score, and median negative contribution score to the TP evaluated pairs, as well as their abundance, prevalence, and stability across all samples.

Table 6.8: Top 10 taxa ranked by median absolute contribution score to the TP evaluated pairs in the anonymised dataset. The median absolute contribution scores $|s|$ are multiplied by 10^3 for presentation.

	Median $\lvert s \rvert$				
Feature	$s > 0$	$s < 0$	Abundance	Prevalence	Stability
Faecalibacterium prausnitzii	10.695	–	3.129	0.973	0.944
Agathobacter rectalis	10.663	–	3.008	0.933	0.853
Ruminococcus bromii	10.652	0.425	2.098	0.703	0.695
Bacteroides uniformis	9.869	–	3.153	0.958	0.928
Alistipes putredinis	9.630	0.425	2.339	0.781	0.904
Eubacterium siraeum	9.617	–	1.541	0.598	0.635
Phocaeicola vulgatus	9.617	–	2.682	0.825	0.874
Roseburia faecis	9.036	0.153	2.061	0.767	0.710
Alistipes shahii	9.036	–	1.951	0.769	0.906
GGB9176 SGB14114	9.036	–	0.707	0.285	0.670

With the exception of *GGB9176 SGB14114*, 9 out of 10 taxa listed in the table are identified in linkage error evaluation as frequently influential features (see Figure 6.5). In particular, Faecalibacterium prausnitzii, Agathobacter rectalis, Bacteroides uniformis, and Phocaeicola vulgatus are among the taxa with the highest positive contributions and also show high abundance, prevalence, and stability. Alistipes putredinis and Alistipes shahii are less abundant and less prevalent but also show high stability. This suggests

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

that the 18 TPs observed for the anonymised dataset are driven by taxa that are frequently influential in the original dataset. Since most of these taxa exhibit distinctive characteristics, their statistical properties may be better preserved after anonymisation, making successful linkage still possible.

Future work could investigate more targeted anonymisation of frequently influential features. Applying stronger anonymisation to these features may help reduce successful linkage while preserving the statistical properties of other features. However, the effects on both privacy protection and analytical utility would need to be evaluated.

In conclusion, although the number of TPs is drastically reduced for the anonymised dataset, correct matches still occur. When some identified pairs are likely to be true, because they are frequently influenced by features derived from demographic factors in the data source, k-anonymity still prevents the identification of any individual among the other $k-1$ individuals in the anonymised group. Regarding the synthetic dataset, successful linkage is impossible, as demonstrated not only by the linkage outcomes but also by the absence of a one-to-one correspondence with the original individuals. Overall, both μ-ANT and synthpop demonstrate strong privacy-protection effectiveness.

Statistical fidelity

Reconstruction quality of the ES decomposition is reported in Table 6.9. For the independent decomposition (rows 1–3), the anonymised dataset achieves the highest values in both explained variance and cosine similarity. For the decomposition using a fixed ES basis (rows 4–5), the anonymised dataset attains higher explained variance and lower reconstruction error than the synthetic dataset. Overall, compared to the baseline explained variance of 0.721 reported in the original study, the ES decomposition demonstrates a satisfactory fit to our datasets.

Table 6.9: Reconstruction quality of the ES decomposition.

(V) feature×sample	(W) feature×ES	Explained variance	Cosine similarity	Reconstruction error
original	original	0.761	0.872	–
anonymised	anonymised	0.915	0.956	–
synthetic	synthetic	0.748	0.865	–
anonymised	original	0.912	–	1.906
synthetic	original	0.744	–	4.375

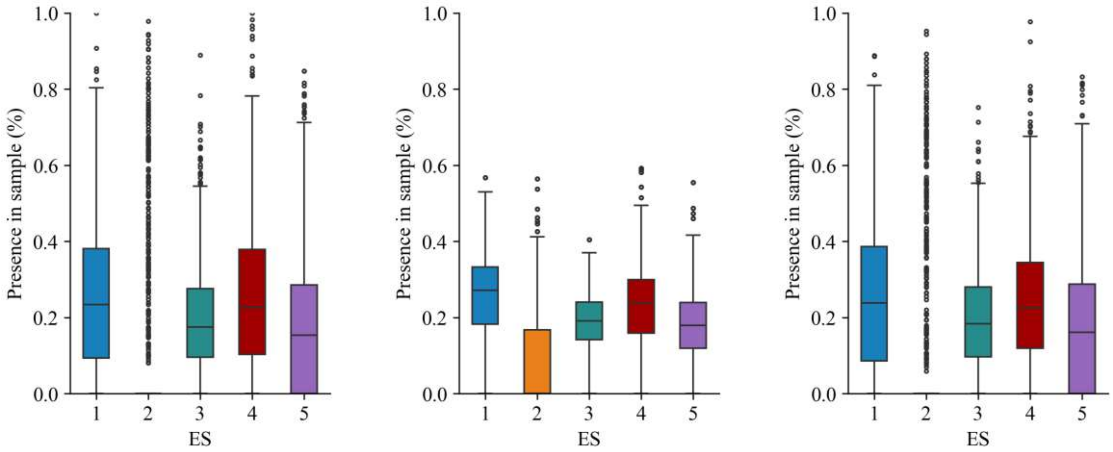
When performing the independent decomposition of the original, anonymised, and synthetic datasets, the resulting ES compositions are shown in Figures 6.7a to 6.7c. Each figure contains five blocks representing ESs. Within each ES, genus-level feature weights are displayed as bar plots, colour-coded by phylum, with a dashed line marking a 25% relative threshold. Pie charts summarise aggregated phylum-level feature weight

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

proportions. This visualisation enables comparison of feature weight consistency across taxonomic ranks.

At the phylum level, the original dataset is dominated by Bacteroidota and Firmicutes, with dominance shifting from the former to the latter from ES1 to ES5. The anonymised and synthetic datasets capture the dominant taxa of the original dataset but differ in the feature-weight proportions and are more similar to each other. At the genus level, the original dataset shows no repeated dominant taxa across ESs, forming distinct separation patterns. Several Bacteroidota taxa, including Bacteroides, Segatella, Phocaeicola, and Alistipes, dominate different ESs, whereas more Firmicutes taxa tend to co-dominate within ESs. The dominant taxa in the synthetic and anonymised datasets are similar to those in the original dataset; however, separation patterns across ESs are better preserved in the synthetic dataset and less so in the anonymised dataset.

For the decomposition using a fixed ES basis derived from the original dataset, Figures 6.6a to 6.6c show the ES distributions across samples in the original, anonymised, and synthetic datasets. The synthetic dataset closely reproduces the distribution observed in the original dataset, whereas the anonymised dataset shows similar medians but does not fully capture the variability.



(a) Original dataset. (b) anonymised dataset. (c) Synthetic dataset.

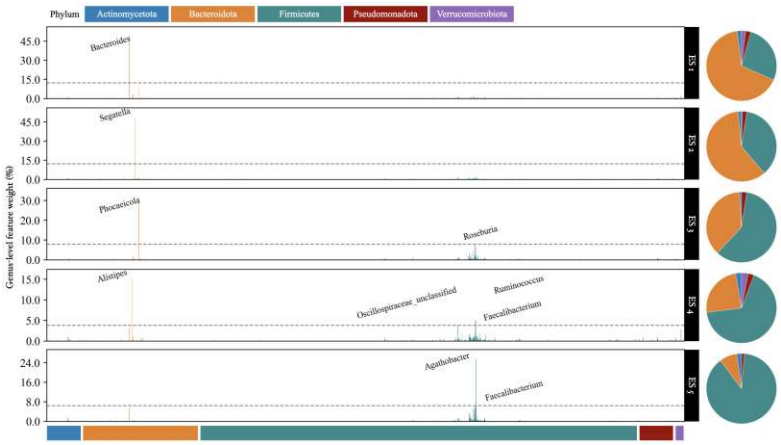
Figure 6.6: ES distribution across samples.

The original dataset exhibits high structural clarity and taxonomic specificity, particularly in the distinct separation patterns of ES composition. Differences in dominant taxa, feature weight proportions, and ES distributions are expected in the anonymised and synthetic datasets, reflecting underlying distributional changes. The synthetic dataset shows some distortion in genus-level dominant taxa but retains clear separation patterns and closely matches the ES distribution. It preserves the underlying compositional structure and variability of the original data, thereby indicating strong statistical fidelity. The anonymised dataset exhibits reduced complexity, making it more amenable to NMF and resulting in high reconstruction quality. However, smoothing effects may lead to a loss

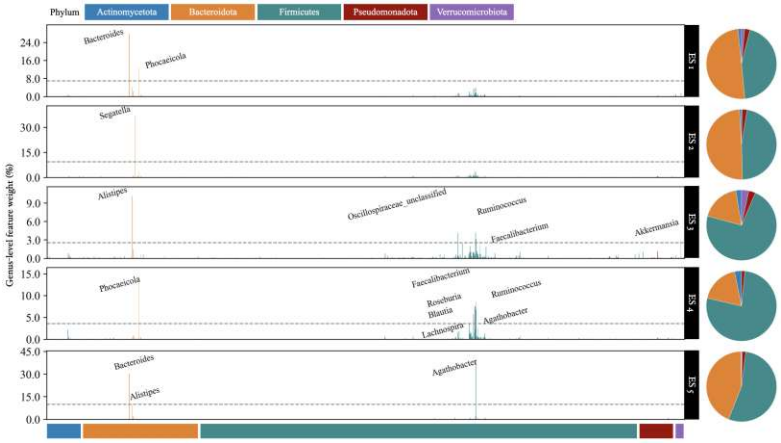
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

of fine-grained structure in ES compositions and ES distributions, while the underlying compositional structure is preserved with reduced variability.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.



(a) Original dataset.



(b) anonymised dataset.

(c) Synthetic dataset.

Figure 6.7: ES composition

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

CHAPTER 7

Conclusion

In this thesis, we evaluate the performance of PMI approaches in settings more challenging than those considered in previous studies, on a dataset with a more complex composition than the ones used in earlier evaluations. These settings include short-, intermediate-, and long-term intervals, as well as true-negative cases, allowing us to examine whether PMI remains feasible in scenarios that more closely reflect real-world applications. We then use the PMI evaluation results to investigate patterns underlying the approaches and their linkage outcomes, including true positives, false positives, true negatives, and false negatives. A substantial overlap between true-positive and false-positive matches leads us to propose an improvement strategy based on mutual nearest neighbours. Finally, we investigate privacy-preserving techniques, including data anonymisation and synthetisation, and evaluate their effects on PMI performance and statistical fidelity.

Below, we summarise the answers to the research questions and outline directions for future work.

7.1 Research Questions

1. **To what extent are personal microbiome identification (PMI) approaches successful as record linkage techniques, based on metrics such as precision, recall, and F1-score, when applied in more challenging PMI settings than those considered in previous studies, including:**
   This question is addressed in Section 6.1. More generally, when precision is the primary concern, such as in matching a single sample against a database, GePMI may be the more suitable choice. When matching between two databases and prioritising overall performance, the nearest-neighbour method appears to be the most suitable approach.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

a) **A more homogeneous study population,**
For one-to-one matching over the short-term interval, both the metagenomic-code method (MetaC) and GePMI achieved 100% precision, whereas the nearest-neighbour method (NearN) showed only slightly lower precision. In contrast, NearN achieved substantially higher recall than the other two approaches. Overall, NearN attained the highest average F1 score (0.99), followed by MetaC (0.95) and GePMI (0.93), suggesting that all three approaches performed well in this setting.

b) **Greater longitudinal diversity, and**
At the intermediate-term interval, the F1 score decreased noticeably, with MetaC achieving 0.71, followed by GePMI at 0.80 and NearN at 0.97. This suggests that PMI performance declines as the time interval between samples increases.

c) **A higher number of true negatives?**
The change is minimal in short-term intervals and larger in intermediate-term intervals. For the nearest-neighbour method, the average F1-score in intermediate-term intervals decreases from 0.98 under paired samples to 0.94 under all samples, while remaining high despite around 38% true negatives.

2. **Which patterns associated with linkage errors can be identified as potentially useful for guiding improvements to PMI approaches in terms of:**

a) **Sample interactions, and**
We found a 57% overlap between true positives and false positives. When a known sample is identified as the nearest neighbour of multiple query samples, it remains the nearest neighbour of its true query sample. When a query sample's nearest known sample is the true match of another query sample, its subsequent nearest known sample is not necessarily the true match. The overlap between TP and FP cases suggests that excluding such cases could improve performance. Accordingly, we incorporated query samples into the dynamic thresholding procedure, which resulted in improved PMI performance.

b) **Feature contributions?**
our analysis suggests that similarity is driven by a small number of features with large positive contributions, whereas dissimilarity is driven by a larger number of features with smaller negative contributions. In addition, we identified the species contributing most strongly to similarity, namely *Bacteroides uniformis*, *Agathobacter rectalis*, *Faecalibacterium prausnitzii*, and *Phocaeicola vulgatus*, together with the frequently dominant features associated with each linkage outcome.

3. **To what extent are data anonymisation and data synthetisation techniques successful as safeguards against personal microbiome identification (PMI) in terms of protecting microbiome data while maintaining statistical fidelity?**

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Both data anonymisation and data synthetisation show strong privacy protection effectiveness and statistical fidelity. Synthetic data preserves the underlying compositional structure and variability of the original data, thereby indicating strong statistical fidelity. Smoothing effects of anonymised data may lead to a loss of fine-grained structure in ES compositions and ES distributions, while the underlying compositional structure is preserved with reduced variability.

7.2 Future work

Several directions for future work can be identified regarding datasets and experiments.

In terms of dataset characteristics, although we identify a dataset that is more challenging than those used in previous studies, access to datasets with larger sample sizes and longer intervals between sampling (more than five years) would be highly valuable.

Moreover, while this thesis uses relative abundance data for PMI evaluation, absolute abundance could be explored in future work. It may provide a more direct representation of changes in individual taxa and potentially preserve person-specific microbial patterns more accurately. However, its benefits for PMI would need to be empirically evaluated.

Regarding PMI evaluation, this thesis conducts the analyses using default settings for each approach. Future work could include sensitivity analyses to examine how the results change under alternative parameter settings. For example, it would be informative to assess how the approaches perform when using other feature representations, such as OTUs, Markers, and kbwindows.. In addition, MetaC's similarity cut-off and GePMI's similarity calculation are based on distance metrics., it would be valuable to examine whether using the same distance measure yields similar results.

Regarding linkage error evaluation, we consider only method-related factors because host-related metadata, such as antibiotic use, probiotic use, and dietary factors, are unavailable. Access to such metadata would enable future work to extend the analysis to host-related factors as well.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Overview of Generative AI Tools Used

- Grammarly is used for grammar checking.
- ChatGPT is used for code generation and debugging.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Figures

2.1 Microbiome collection and analysis workflow. The stages include (i) sample collection (top-left), (ii) experimental procedure (top-middle), (iii) bioinformatics pipeline (bottom-right), and (iv) analysis (bottom-left). Reproduced from [KLFP24], licensed under CC BY 4.0. 7

4.1 Timelines and sampling time points of the HPFS cohort, the MLVS cohort, and the MLVS faecal microbiome dataset. n = number of participants. . . 21
4.2 Stratification of the mixed-rank profile into rank-specific profiles. The mixed-rank feature abundance profile shown in Figure 4.2 was stratified into eight rank-specific profiles, each containing features from exactly one taxonomic rank, with two example features shown for each rank. 24
4.3 Preprocessing workflow and number of samples by step. n = number of samples. 25

5.1 Matrix factorisation for ES decomposition. 44

6.1 PMI linkage outcome proportions by approach on paired samples. 49
6.2 PMI linkage outcome proportions by approach on all samples. 50
6.3 Feature abundance, prevalence, and stability 59
6.4 Overall feature contribution patterns 61
6.5 Frequently influential features by linkage outcome. 62
6.6 ES distribution across samples. 66
6.7 ES composition 68

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Tables

2.1 Raw microbiome data types. Adapted from [KLFP24], licensed under CC BY 4.0. Capabilities*: Resolution, the ability to distinguish between different microbes or genes; Selectivity, the ability to target specific microbes or genes for analysis; Capacity, the amount of data that can be generated and analysed. 8

3.1 Detection thresholds in the metagenomic-code method. Adapted from [FHM+15]. Features with abundance between the confident detection and confident non-detection thresholds are considered ambiguous. 10
3.2 Feature abundance limits in the nearest-neighbour method. Reproduced from [MHE24]. 12
3.3 Metagenomic feature types in the Franzosa dataset. Adapted from [FHM+15]. Units: RPKM = reads per kilobase per million sample reads. 13
3.4 Contingency matrix of McNemar's test. PMI approaches: MetaC = the metagenomic-code method; NearN = the nearest-neighbour method. . . . 15
3.5 Mean precision, recall, and F1-score over metagenomic feature types for the metagenomic-code and nearest-neighbour methods. Reproduced from [MHE24]. PMI approaches: MetaC = the metagenomic-code method; NearN = the nearest-neighbour method. 16
3.6 Summary statistics of the multi-cohort dataset. Visits: the number of visits (number of participants with the corresponding visit count). 17

4.1 Comparison of the Franzosa dataset, the multi-cohort dataset, and the MLVS faecal microbiome dataset with respect to study population homogeneity, longitudinal diversity, and the number of true negatives. "?" = not reported in the original source. Samples by time point: time point (number of samples with the corresponding time point). 22
4.2 Excerpt of *Species* features, for three subjects (columns 2–4). 23
4.3 Number of detected features by taxonomic rank. 25

5.1 Visit combinations. 28
5.2 Summary statistics of paired samples by visit pair (one-to-one matching). 28
5.3 Summary statistics of all samples by visit (matching including true negatives). 29

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

5.4 Evaluation scenarios and relevant summary statistic. Matching mode: 1:1 = one-to-one matching; +TN = matching including true negatives. Time interval: ST = short-term; IT = intermediate-term. 29
5.5 Input data structure of Cochran's Q test. 34
5.6 Dataset variant assignment for PMI. 43
5.7 Dataset variant assignment for ES composition. 44
5.8 Evaluation of the number of ESs. 45

6.1 PMI effectiveness by distance measure for GePMI. For each visit combination, the highest F1-score among distance measures is highlighted in green, the second-highest in blue, and the lowest in red. 48
6.2 Average PMI effectiveness by approach. For each evaluation scenario, the highest value among PMI approaches is highlighted in green. Matching mode: 1:1 = one-to-one matching; +TN = matching including true negatives. Time interval: ST = short-term; IT = intermediate-term. PMI Approach: MetaC = the metagenomic-code method; NearN = the nearest-neighbour method. 51
6.3 PMI effectiveness by approach. For each matching mode and visit combination, the highest F1-score among PMI approaches is highlighted in green and the lowest in red. 52
6.4 Summary statistics of TP–FP proposal conflicts. For each evaluation scenario, conflict rates above 50% are highlighted in red. Matching mode: 1:1 = one-to-one matching; +TN = matching including true negatives. Visit combination: time point of known sample to time point of query sample. 53
6.5 PMI effectiveness by improvement strategy for the nearest-neighbour method. For each evaluation scenario, the highest TP and TN counts among improvement strategies are highlighted in blue. Matching mode: 1:1 = one-to-one matching; +TN = matching including true negatives. Visit combination: time point of known sample to time point of query sample. Time interval: ST = short-term; IT = intermediate-term. Improvement strategy: NearN = the original nearest-neighbour method; MNN = Mutual nearest neighbour; MNN-RM = Mutual nearest neighbour with rematching. 58
6.6 Summary statistics of feature abundance, prevalence, and stability. 59
6.7 Summary statistics of overall feature contribution by linkage outcome. The values are reported in thousandths. 60
6.8 Top 10 taxa ranked by median absolute contribution score to the TP evaluated pairs in the anonymised dataset. The median absolute contribution scores $|s|$ are multiplied by 10^3 for presentation. 64
6.9 Reconstruction quality of the ES decomposition. 65

1 Samples removed during the preprocessing workflow. 83
2 Upper-tail critical values of the χ^2 distribution. 84

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

3 Statistical tests of PMI effectiveness by approach. Among all p-values, failure to reject the null hypothesis is highlighted by gray shading, while rejection of the null hypothesis ($p < 0.05$) is highlighted by red shading 85
4 PMI effectiveness by approach on the anonymised dataset. 86
5 PMI effectiveness by approach on the synthetic dataset. 87
6 Feature abundance profile. 88

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Algorithms

6.1 Nearest-neighbour PMI 54
6.2 Nearest-neighbour PMI: MNN without rematching 55
6.3 Nearest-neighbour PMI: MNN with Rematching 56

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Appendix

Table 1: Samples removed during the preprocessing workflow.

Reason	Samples
Failed during processing	SRR5155754, SRR5165642, SRR5165653, SRR5165688, SRR5165702, SRR5165721, SRR5165892, SRR5165974
Invalid annotations	SRR5165793, SRR5165803

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 2: Upper-tail critical values of the χ^2 distribution.

	Significance level α									
Degrees of freedom	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	–	–	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

TU WIEN Bibliothek Your knowledge hub
Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 3: Statistical tests of PMI effectiveness by approach. Among all p-values, failure to reject the null hypothesis is highlighted by gray shading, while rejection of the null hypothesis ($p < 0.05$) is highlighted by red shading.

Exp. scenario		Cochran's Q test		McNemar's test								
		MC–GE–NN		MC–GE			MC–NN			GE–NN		
Matching	Visit	Q	p	χ^2	p	p_{adj}	χ^2	p	p_{adj}	χ^2	p	p_{adj}
paired	v1:v2	34.667	<0.001	6.618	0.010	0.010	12.071	<0.001	0.001	28.033	<0.001	<0.001
paired	v2:v3	70.977	<0.001	4.516	0.034	0.034	58.141	<0.001	<0.001	42.023	<0.001	<0.001
paired	v3:v4	21.941	<0.001	0.346	0.556	0.556	16.000	<0.001	<0.001	15.059	<0.001	<0.001
all	v1:v2	30.780	<0.001	5.600	0.018	0.018	10.562	0.001	0.002	25.290	<0.001	<0.001
all	v1:v3	53.255	<0.001	10.776	0.001	0.001	40.477	<0.001	<0.001	18.150	<0.001	<0.001
all	v1:v4	63.047	<0.001	18.015	<0.001	<0.001	48.944	<0.001	<0.001	15.789	<0.001	<0.001
all	v2:v3	68.780	<0.001	8.862	0.003	0.003	53.778	<0.001	<0.001	31.130	<0.001	<0.001
all	v2:v4	65.453	<0.001	13.653	<0.001	<0.001	53.440	<0.001	<0.001	21.811	<0.001	<0.001
all	v3:v4	19.500	<0.001	0.893	0.345	0.345	14.815	<0.001	<0.001	11.529	<0.001	0.001

Table 4: PMI effectiveness by approach on the anonymised dataset.

Mode	Visit	App.	Prec.	Rec.	F1	TP	FP	TP+FP	TN	FN
1:1	v1:v2 (ST)	MetaC	0.00	0.00	0.00	0	0	0	–	200
		GePMI	0.00	0.00	0.00	0	2	1	–	197
		NearN	0.14	0.01	0.02	2	12	–	–	186
1:1	v2:v3 (IT)	MetaC	0.00	0.00	0.00	0	0	0	–	156
		GePMI	0.50	0.01	0.01	1	1	0	–	154
		NearN	0.22	0.01	0.03	2	7	–	–	147
1:1	v3:v4 (ST)	MetaC	0.00	0.00	0.00	0	0	0	–	244
		GePMI	0.25	0.00	0.01	1	3	0	–	240
		NearN	0.36	0.02	0.04	5	9	–	–	230
+TN	v1:v2 (ST)	MetaC	0.00	0.00	0.00	0	0	0	7	200
		GePMI	0.00	0.00	0.00	0	2	1	7	197
		NearN	0.13	0.01	0.02	2	13	–	7	185
+TN	v1:v3 (IT)	MetaC	0.00	0.00	0.00	0	0	0	95	159
		GePMI	0.00	0.00	0.00	0	3	0	93	158
		NearN	0.00	0.00	0.00	0	23	-	83	148
+TN	v1:v4 (IT)	MetaC	0.00	0.00	0.00	0	0	0	91	158
		GePMI	0.00	0.00	0.00	0	2	0	90	157
		NearN	0.00	0.00	0.00	0	24	-	81	144
+TN	v2:v3 (IT)	MetaC	0.00	0.00	0.00	0	0	0	98	156
		GePMI	0.00	0.00	0.00	0	4	0	95	155
		NearN	0.04	0.01	0.01	1	23	–	82	148
+TN	v2:v4 (IT)	MetaC	0.00	0.00	0.00	0	0	0	94	155
		GePMI	0.00	0.00	0.00	0	4	0	93	152
		NearN	0.03	0.01	0.01	1	31	–	75	142
+TN	v3:v4 (ST)	MetaC	0.00	0.00	0.00	0	0	0	5	244
		GePMI	0.25	0.00	0.01	1	3	0	5	240
		NearN	0.31	0.02	0.04	5	11	–	5	228

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 5: PMI effectiveness by approach on the synthetic dataset.

Mode	Visit	App.	Prec.	Rec.	F1	TP	FP	TP+FP	TN	FN
1:1	v1:v2 (ST)	MetaC	0.03	0.01	0.01	1	33	0	–	166
		GePMI	0.00	0.00	0.00	0	0	0	–	200
		NearN	0.00	0.00	0.00	0	13	–	–	187
1:1	v2:v3 (IT)	MetaC	0.00	0.00	0.00	0	18	0	–	138
		GePMI	0.00	0.00	0.00	0	0	0	–	156
		NearN	0.00	0.00	0.00	0	10	–	–	146
1:1	v3:v4 (ST)	MetaC	0.00	0.00	0.00	0	24	0	–	220
		GePMI	0.00	0.00	0.00	0	0	0	–	244
		NearN	0.00	0.00	0.00	0	17	–	–	227
+TN	v1:v2 (ST)	MetaC	0.00	0.00	0.00	0	35	1	5	166
		GePMI	0.00	0.00	0.00	0	0	0	7	200
		NearN	0.00	0.00	0.00	0	14	–	7	186
+TN	v1:v3 (IT)	MetaC	0.00	0.00	0.00	0	46	0	73	135
		GePMI	0.00	0.00	0.00	0	0	0	95	159
		NearN	0.00	0.00	0.00	0	11	–	92	151
+TN	v1:v4 (IT)	MetaC	0.00	0.00	0.00	0	43	0	72	134
		GePMI	0.00	0.00	0.00	0	0	0	91	158
		NearN	0.00	0.00	0.00	0	11	–	90	148
+TN	v2:v3 (IT)	MetaC	0.00	0.00	0.00	0	28	0	88	138
		GePMI	0.00	0.00	0.00	0	0	0	98	156
		NearN	0.00	0.00	0.00	0	16	–	92	146
+TN	v2:v4 (IT)	MetaC	0.00	0.00	0.00	0	28	0	83	138
		GePMI	0.00	0.00	0.00	0	0	0	94	155
		NearN	0.00	0.00	0.00	0	18	–	88	143
+TN	v3:v4 (ST)	MetaC	0.00	0.00	0.00	0	30	0	5	214
		GePMI	0.00	0.00	0.00	0	0	0	5	244
		NearN	0.00	0.00	0.00	0	17	–	5	227

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Table 6: Feature abundance profile.

	sample			
taxa	1	2	...	n
1	x_{11}	x_{12}	...	x_{1n}
2	x_{21}	x_{22}	...	x_{2n}
⋮	⋮	⋮	⋮	⋮
m	x_{m1}	x_{m2}	...	x_{mn}
Total	1	1	...	1

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Bibliography

[AKP+23] Nikhil Aggarwal, Shohei Kitano, Ginette Ru Ying Puah, Sandra Kittelmann, In Young Hwang, and Matthew Wook Chang. Microbiome and Human Health: Current Understanding, Engineering, and Enabling Technologies. *Chemical Reviews*, 123(1):31–72, January 2023.

[BJS+10] Megan A Bohensky, Damien Jolley, Vijaya Sundararajan, Sue Evans, David V Pilcher, Ian Scott, and Caroline A Brand. Data Linkage: A powerful research tool with potential problems. *BMC Health Services Research*, 10(1):346, December 2010.

[BRF+20] Gabriele Berg, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, Luca Cocolin, Kellye Eversole, Gema Herrero Corral, Maria Kazou, Linda Kinkel, Lene Lange, Nelson Lima, Alexander Loy, James A. Macklin, Emmanuelle Maguin, Tim Mauchline, Ryan McClure, Birgit Mitter, Matthew Ryan, Inga Sarand, Hauke Smidt, Bettina Schelkle, Hugo Roume, G. Seghal Kiran, Joseph Selvin, Rafael Soares Correa De Souza, Leo Van Overbeek, Brajesh K. Singh, Michael Wagner, Aaron Walsh, Angela Sessitsch, and Michael Schloter. Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1):103, December 2020.

[BY01] Yoav Benjamini and Daniel Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[Coc50] W. G. Cochran. The Comparison of Percentages in Matched Samples. *Biometrika*, 37(3/4):256, December 1950.

[CVL+24] Urte Clausen, Sören-Tobias Vital, Pia Lambertus, Martina Gehler, Sabine Scheve, Lars Wöhlbrand, and Ralf Rabus. Catabolic Network of the Fermentative Gut Bacterium *Phocaeicola vulgatus* (Phylum Bacteroidota) from a Physiologic-Proteomic Perspective. *Microbial Physiology*, 34(1):88–107, 2024.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[DRQ+16] Sylvia H. Duncan, Wendy R. Russell, Andrea Quartieri, Maddalena Rossi, Julian Parkhill, Alan W. Walker, and Harry J. Flint. Wheat bran promotes enrichment within the human colonic microbiota of butyrate-producing bacteria that release ferulic acid. *Environmental Microbiology*, 18(7):2214–2225, July 2016.

[FAO+23] Clemence Frioux, Rebecca Ansorge, Ezgi Ozkurt, Chabname Ghassemi Nedjad, Joachim Fritscher, Christopher Quince, Sebastian M. Waszak, and Falk Hildebrand. Enterosignatures define common bacterial guilds in the human gut microbiome. *Cell Host & Microbe*, 31(7):1111–1125.e6, July 2023.

[FHM+15] Eric A. Franzosa, Katherine Huang, James F. Meadow, Dirk Gevers, Katherine P. Lemon, Brendan J. M. Bohannan, and Curtis Huttenhower. Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences*, 112(22), June 2015.

[GBVR03] Lifang Gu, Rohan Baxter, Deanne Vickers, and Chris Rainsford. Record Linkage: Current Practice and Future Directions. CMIS Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, ACT, Australia, 2003.

[HDB+17] Katie Harron, Chris Dibben, James Boyd, Anders Hjern, Mahmoud Azimaee, Mauricio L. Barreto, and Harvey Goldstein. Challenges in administrative data linkage for research. *Big Data & Society*, 4(2):2053951717745678, December 2017.

[HLMM18] Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, May 2018.

[HME22] Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. Utility and Privacy Assessment of Synthetic Microbiome Data. In *Data and Applications Security and Privacy XXXVI*, volume 13383, pages 15–27. Springer International Publishing, Cham, 2022. Series Title: Lecture Notes in Computer Science.

[Hol79] Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[HRW+21] Pranvera Hiseni, Knut Rudi, Robert C. Wilson, Finn Terje Hegge, and Lars Snipen. HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. *Microbiome*, 9(1):165, December 2021.

[Hum12] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[KLFP24] Bablu Kumar, Erika Lorusso, Bruno Fosso, and Graziano Pesole. A comprehensive overview of microbiome data in the light of machine learning applications: categorization, accessibility, and future directions. *Frontiers in Microbiology*, 15:1343572, February 2024.

[LSG+12] Catherine A. Lozupone, Jesse I. Stombaugh, Jeffrey I. Gordon, Janet K. Jansson, and Rob Knight. Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220–230, September 2012.

[LSL+19] Xiaosun Liu, Li Shao, Xia Liu, Feng Ji, Ying Mei, Yiwen Cheng, Fengping Liu, Chongxian Yan, Lanjuan Li, and Zongxin Ling. Alterations of gastric mucosal microbiota across different stomach microhabitats in a cohort of 276 patients with gastric cancer. *EBioMedicine*, 40:336–348, February 2019.

[MAAD+18] Raaj S. Mehta, Galeb S. Abu-Ali, David A. Drew, Jason Lloyd-Price, Ayshwarya Subramanian, Paul Lochhead, Amit D. Joshi, Kerry L. Ivey, Hamed Khalili, Gordon T. Brown, Casey DuLong, Mingyang Song, Long H. Nguyen, Himel Mallick, Eric B. Rimm, Jacques Izard, Curtis Huttenhower, and Andrew T. Chan. Stability of the human faecal microbiome in a cohort of adult men. *Nature Microbiology*, 3(3):347–355, January 2018.

[McN47] Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, June 1947.

[MHE24] Rudolf Mayer, Markus Hittmeir, and Andreas Ekelhart. Distance-based linkage of personal microbiome records for identification and its privacy implications. *Computers & Security*, 136:103538, January 2024.

[MJS+14] Kathleen Machiels, Marie Joossens, João Sabino, Vicky De Preter, Ingrid Arijs, Venessa Eeckhaut, Vera Ballet, Karolien Claes, Filip Van Immerseel, Kristin Verbeke, Marc Ferrante, Jan Verhaegen, Paul Rutgeerts, and Séverine Vermeire. A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut*, 63(8):1275–1283, August 2014.

[MKH23] Rudolf Mayer, Alicja Karlowicz, and Markus Hittmeir. k-Anonymity on Metagenomic Features in Microbiome Databases. In *Proceedings of the 18th International Conference on Availability, Reliability and Security*, pages 1–11, Benevento Italy, August 2023. ACM.

[MMR+13] S Miquel, R Martín, O Rossi, Lg Bermúdez-Humarán, Jm Chatel, H Sokol, M Thomas, Jm Wells, and P Langella. Faecalibacterium prausnitzii and human intestinal health. *Current Opinion in Microbiology*, 16(3):255–261, June 2013.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[MSW+13] Pradeep Reddy Marri, Debra A. Stern, Anne L. Wright, Dean Billheimer, and Fernando D. Martinez. Asthma-associated differences in microbial composition of induced sputum. *Journal of Allergy and Clinical Immunology*, 131(2):346–352.e3, February 2013.

[NRD16] Beata Nowok, Gillian M. Raab, and Chris Dibben. synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74(11):1 – 26, October 2016.

[PDLDDT25] Cecile Melisse Ponce De Leon-Derecho and Genevieve Dable-Tupas. Ethical considerations in microbiome research. In *Human Microbiome Drug Targets*, pages 179–188. Elsevier, 2025.

[Rei05] Jerome P Reiter. Estimating Risks of Identification Disclosure in Microdata. *Journal of the American Statistical Association*, 100(472):1103–1112, December 2005.

[RKS+16] Jaime A. Rosero, Jirˇí Killer, Hana Sechovcová, Jakub Mrázek, Oldrˇich Benada, Katerˇina Fliegerová, Jaroslav Havlík, and Jan Kopečný. Reclassification of Eubacterium rectale (Hauduroy et al. 1937) Prévot 1938 in a new genus Agathobacter gen. nov. as Agathobacter rectalis comb. nov., and description of Agathobacter ruminis sp. nov., isolated from the rumen contents of sheep and cows. *International Journal of Systematic and Evolutionary Microbiology*, 66(2):768–773, February 2016.

[Sam01] P. Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, December 2001.

[SLT+22] Paul O. Sheridan, Petra Louis, Eleni Tsompanidou, Sophie Shaw, Hermie J. Harmsen, Sylvia H. Duncan, Harry J. Flint, and Alan W. Walker. Distribution, organization and expression of genes concerned with anaerobic lactate utilization in human intestinal bacteria. *Microbial Genomics*, 8(1), January 2022.

[SMDF+20] David Sánchez, Sergio Martínez, Josep Domingo-Ferrer, Jordi Soria-Comas, and Montserrat Batet. **μ** *-ANT* : semantic microaggregation-based anonymization tool. *Bioinformatics*, 36(5):1652–1653, March 2020.

[SP19] Brandon H. Schlomann and Raghuveer Parthasarathy. Timescales of gut microbiome dynamics. *Current Opinion in Microbiology*, 50:56–63, August 2019.

[SPJB02] Rainer Simmering, Holger Pforte, Gisela Jacobasch, and Michael Blaut. The growth of the flavonoid-degrading intestinal bacterium, Eubacterium ramulus, is stimulated by dietary flavonoids in vivo. *FEMS Microbiology Ecology*, 40(3):243–248, June 2002.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN

[SPW+08] Harry Sokol, Bénédicte Pigneur, Laurie Watterlot, Omar Lakhdari, Luis G. Bermúdez-Humarán, Jean-Jacques Gratadoux, Sébastien Blugeon, Chantal Bridonneau, Jean-Pierre Furet, Gérard Corthier, Corinne Grangette, Nadia Vasquez, Philippe Pochart, Germain Trugnan, Ginette Thomas, Hervé M. Blottière, Joël Doré, Philippe Marteau, Philippe Seksik, and Philippe Langella. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proceedings of the National Academy of Sciences*, 105(43):16731–16736, October 2008.

[UMPK12] Luke K Ursell, Jessica L Metcalf, Laura Wegener Parfrey, and Rob Knight. Defining the human microbiome. *Nutrition Reviews*, 70:S38–S44, August 2012.

[WDH+10] Ben P. Willing, Johan Dicksved, Jonas Halfvarson, Anders F. Andersson, Marianna Lucio, Zongli Zheng, Gunnar Järnerot, Curt Tysk, Janet K. Jansson, and Lars Engstrand. A Pyrosequencing Study in Twins Shows That Gastrointestinal Microbial Profiles Vary With Inflammatory Bowel Disease Phenotypes. *Gastroenterology*, 139(6):1844–1854.e1, December 2010.

[WLW+18] Zicheng Wang, Huazhe Lou, Ying Wang, Ron Shamir, Rui Jiang, and Ting Chen. GePMI: A statistical model for personal intestinal microbiome identification. *npj Biofilms and Microbiomes*, 4(1):20, September 2018.

[ZNS+23] Ruwen Zhou, Siu Kin Ng, Joseph Jao Yiu Sung, Wilson Wen Bin Goh, and Sunny Hei Wong. Data pre-processing for analyzing microbiome data – A mini review. *Computational and Structural Biotechnology Journal*, 21:4804–4815, January 2023.

Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.
TU Bibliothek Your knowledge hub WIEN