Analysis of Human Gut Metagenomes for the Prediction of Host Traits with Tree Ensemble Machine Learning Models

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Albane Mialy Ruaud
aus Paris / Frankreich

Tübingen
2021

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	28.06.2021
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Ruth E. Ley
2. Berichterstatter:	Prof. Dr. Daniel Huson
3. Berichterstatter:	Dr. Richard Neher

Abstract

The human gut microbiota is made of a myriad of microorganisms, among which not only bacteria but also archaea. Present at lower abundances, technically more challenging to quantify, and under-represented in databases, archaea are often overseen when describing the human gut microbiome. Nonetheless, the main archaeon in terms of prevalence and abundance is *Methanobrevibacter smithii*, family *Methanobacteriaceae*. It has been associated with various host phenotypes such as slow transit or diet habits. Remarkably, contrasting evidence shows an association between *M. smithii* and body mass index (BMI): it is enriched in lean or obese individuals according to population studies. Reasonable hypotheses relying on the metabolism of the archaeon support these conflicting findings. For instance, its slow replication time supports its association with slow transit.

M. smithii and all members of the *Methanobacteriaceae* family are methanogens: their metabolism relies on the reduction of simple carbon molecules to methane. In the human gut, methanogenesis starts from bacterial fermentation products. In particular, H_2 and CO_2 are the primary substrates of *M. smithii*, formate can also be used but with a lower energy yield. By uptaking fermentation products, *M. smithii* can boost specific fermentation pathways, consequently affecting the production of short-chain fatty acids (SCFA). These byproducts of bacterial fermentation are absorbed by the host, where they mediate host energy and inflammatory metabolisms. Accordingly, its overall effect may depend on the fermentation potential of the gut microbiome, itself defined by the microbiome composition. Hence, *M. smithii* may influence its host by consuming fermentation products. Because we know so little about the interactions between *M. smithii* and fermenting bacteria, gaining knowledge on their diversity and specificity and the underlying mechanisms would improve our understanding of methanogens' role in the human gut.

This work aims at providing insights into the associations between *M. smithii* and gut bacteria. Due to the fastidiousness of methanogens' culture, I performed a meta-analysis of human gut metagenomes using machine learning models. To decipher the variable interactions captured by the model, I developed a tool for interpreting tree ensemble models. My new method allowed me to infer biologically relevant associations between the methanogen and components of the human gut environment. In particular, I found a clear association between *M. smithii* and an uncultured family of the *Christensenellales* order, as well as members of the *Oscillospirales* order predicted to have a slow replication time and be associated with slow transit. Furthermore, predictions from the model revealed a gradient in relative abundances of a core group of taxa associated with the colonization of human guts by *Methanobacteriaceae*. This gradient generally followed microbiome composition types, i.e., enterotypes, previously correlated with human population traits. This suggests that associations between methanogens and phenotypes known to be associated with certain enterotypes, such as BMI is correlated with the ETB enterotype, may be spurious. Then, I further explored the association between *M. smithii* and members of the *Christensenellales* order. For this, I compared cocultures of *M. smithii* with *Christensenella minuta*, a human gut isolate of the *Christensenellaceae* family, and *Bacteroides thetaiotaomicron*, a common H_2-producer from the human gut. Results demonstrated a syntrophy via H_2-transfer between *Christensenellaceae* and the methanogen, accompanied by a switch in SCFA production.

Altogether, my findings complement the current knowledge on interactions between the human gut methanogen *M. smithii* and fermenting bacteria. They support the hypothesis that *M. smithii* preferentially interacts with specific H_2-producers in the human gut, e.g., members of the *Christensenellales* order, as well as a core group of bacteria favoring its colonization of the gut environment. Syntrophy may underlie the identified associations, with potential effects on bacterial fermentation. In addition, my method for interpreting machine learning models applies to all sorts of problems being studied with tree ensemble models. Thus, its potential in helping understand complex systems is not limited to the microbiome field and will hopefully appear useful to other researchers in the future.

Zusammenfassung

Das Darmmikrobiom des Menschen besteht aus einer Vielzahl von Mikroorganismen, darunter nicht nur Bakterien, sondern auch Archaeen. Archaeen, die in geringerer Häufigkeit vorhanden, technisch schwieriger zu quantifizieren und in Datenbanken unterrepräsentiert sind, werden bei der Beschreibung des menschlichen Darmmikrobioms häufig übergangen. Ungeachtet dessen ist *Methanobrevibacter smithii*, ein Mitglied der Familie der *Methanobacteriaceae*, das Hauptarchäon in Bezug auf Prävalenz und Häufigkeit. Es wurde mit verschiedenen Phänotypen des menschlichen Wirtes wie verringerte Darmbeweglichkeit oder Ernährungsgewohnheiten verbunden. Bemerkenswerterweise zeigen verschiedene Studien einen widersetzlichen Zusammenhang zwischen *M. smithii* und dem Body-Mass-Index (BMI): Während einige Studien eine erhöhte Abundanz von *M. smithii* in schlanken Menschen aufweisen, zeigen andere eine höhere Häufigkeit in fettleibigen Menschen. Begründete Hypothesen, die sich auf den Stoffwechsel des Archäons beruhen, stützen diese widersprüchlichen Befunde. Beispielsweise unterstuezt seine langsame Replikationszeit die Assoziation mit einer verringerten Darmbeweglichkeit.

M. smithii und alle anderen Mitglieder der *Methanobacteriaceae*-Familie sind Methanogene: Ihr Stoffwechsel beruht auf der Reduktion einfacher Kohlenstoffmoleküle zu Methan. Im menschlichen Darms geht die Methanogenese von bakteriellen Fermentationsprodukten aus. Insbesondere H_2 und CO_2 sind die primären Substrate von *M. smithii*. Auch Formiat kann verwendet werden, jedoch mit einer geringeren Energieausbeute. Durch die Absorption von Fermentationsprodukten kann *M. smithii* bestimmte Fermentationsprozesse ankurbeln und folglich die Produktion kurzkettiger Fettsäuren (SCFA) beeinflussen. Solche Nebenprodukte der bakteriellen Fermentation werden vom Wirt absorbiert und beeinflussen seinen Energie- und Entzündungsstoffwechsel. Dementsprechend könnte die Gesamtwirkung von *M. smithii* auf den menschlichen Wirt vom Fermentationspotential des Darmmikrobioms abhängen, das wiederum durch die Mikrobiomzusammensetzung definiert ist. Auf diese Weise kann *M. smithii* im Zusammenspiel mit dem Mikrobiom seinen Wirt durch den Verzehr von Fermentationsprodukten beeinflussen. Weil so wenig über die Wechselwirkungen zwischen *M. smithii* und fermentierenden Bakterien bekannt ist, kann das Wissen über ihre Vielfalt und Spezifität, sowie ihrer zugrunde liegenden Mechanismen unser Verständnis der Rolle der Methanogenen im menschlichen Darm vertiefen.

Ziel dieser Arbeit ist es, Einblicke in die Beziehung zwischen *M. smithii* und Darmbakterien zu geben. Da Methanogenkulturen anspruchsvoll sind, führte ich eine Metaanalyse menschlicher Darmmetagenome mittels Machine-Learning Modellen durch. Um die vom Modell erfassten variablen Interaktionen zu entschlüsseln, habe ich ein Tool zur Interpretation von Tree-Ensemble-Modellen entwickelt. Mit meiner neuen Methode konnte ich biologisch relevante Zusammenhänge zwischen Methanogen und Komponenten des menschlichen Darmmilieus erschließen. Insbesondere fand ich einen klaren Zusammenhang zwischen *M. smithii* und einer nicht kultivierten Familie in der Ordnung der *Christensenellales* sowie Mitgliedern des *Oscillospirales*-Ordnung. Für Letztere wurde ebenfalls eine langsame Replikationszeit prognostiziert. Darüber hinaus zeigten die Prognosen des Modells, dass die relative Häufigkeit einer Kerngruppe von Taxa graduell mit der Besiedlung des menschlichen Darms durch *Methanobacteriaceae* einhergehent. Dieser Gradient folgte im Allgemeinen den Mikrobiom-Zusammensetzung Typen, Enterotypen genannt, die zuvor mit menschlichen Bevölkerungsmerkmalen in Verbindung gebracht wurden. Dies deutet darauf hin, dass Zusammenhänge zwischen Methanogenen und bestimmten Phänotypen, von welchen eine Verbindung mit den Enterotypen bekannt ist, beispielsweise die Korrelation von BMI mit dem ETB-Enterotyp, möglicherweise falsch sind. Weiterhin untersuchte ich die Beziehung zwischen *M. smithii* und den Mitgliedern der Christensenellales-Ordenung. Zu diesem Zweck verglich ich Kokulturen von *M. smithii* mit *Christensenella minuta*, einem menschlichen Darmisolat aus der Familie der *Christensenellaceae*, und *Bacteroides thetaiotaomicron*, einem verbreiteten H_2-Produzenten aus dem menschlichen Darm. Die Ergebnisse zeigten eine Syntrophie über H_2-Transfer zwischen *Christensenellaceae* und dem Methanogen, begleitet von einem Wechsel in der Produktion kurzkettiger Fettsäuren.

Insgesamt kann ich durch meine Ergebnisse das aktuelle Wissen über Wechselwirkungen zwischen dem menschlichen Darmmethanogen *M. smithii* und fermentierenden Bakterien ergänzen. Sie unterstützten die Hypothese, dass *M. smithii* bevorzugt mit spezifischen H_2-Produzenten im menschlichen Darm interagiert, z. B. Mitgliedern der *Christensenellales*-Ordnung, sowie mit einem Kernmikrobiom, das die Besiedlung von *M. smithii* in der Darmumgebung begünstigt. Dem identifizierten Zusammenhang kann eine Syntrophie zugrunde liegen, die potenzielle Auswirkungen auf die bakterielle Fermentation hat. Darüber hinaus lässt sich meine Methode zur Interpretation von Machine-Learning-Modellen auf alle Arten von Problematiken, die mit Tree-Ensemble-Modellen untersucht werden, anwenden. Da diese Methodologie zum Verständnis komplexer Systeme beiträgt und nicht auf das Mikrobiomfeld beschränkt ist, kann es in Zukunft auch Forschern aus anderen Themengebieten von Nutzen sein.

Acknowledgements

First, I would like to thank Ruth E. Ley and Nicholas D. Youngblut for the opportunity they gave me to do my Ph.D. in the Department of Microbiome Science at the MPI-DB. You allowed me to explore the paths I encountered and trusted me to independently perform my research, even though it meant changing the Ph.D. plan and, sometimes, walking in the dark. I also would like to thank the other members of my thesis committee advisory board, Daniel Huson and Richard Neher.

Collectively, you have all followed my progress and digressions, and have provided valuable feedback during the Ph.D. Thank you so much for that.

I am more than grateful to Sofia Esquivel-Elizondo and Niklas Pfister. They supervised me, provided guidance and feedback during the two main projects of my thesis. I loved working with them and am looking forward to future collaborations.

All the lab work I did started with Selma Atan; she taught me everything I know about anaerobic work and has been a great friend. A big thanks as well to the Angenent lab with whom we shared space when I started the Ph.D. and who notably participated to my HPLC, GC, glove box, and gas station education.

I also want to thank Christian Feldhaus and Aurora Penzera from the Light Microscopy Facility, who provided valuable help with microscopy and kept the moods up! Thanks to Andre Noll too, for keeping the cluster tidy when all my jobs were failing or causing problems. In general, all facility and lab staff of the MPI-DB have participated in making my experience smooth; thanks.

A significant part of my work and education relies on open science: publicly available data, open-source software, open access scientific articles, etc. It is thus important to me to thank all who contribute to it; I sincerely hope these practices will become the norm in the future.

Vielen danke to Tanja Schön and Laura Jentsch who translated my abstract, y muchas gracias a Andrea Borbón who did magic on Figure 4.6.

Then, I would like to thank all the friends and colleagues of the Microbiome Science department, who contributed to my work through discussions, lab meetings, and walks to the vending machine: Jacobo de la Cuesta, Leonardo Moreno, Andrea Borbón, Selma Atan, Liam Fitzstevens, Daphne Welter, Stacey Heaver, Tanja Schön, Guillermo Luque, Victor Schmidt, Claudia Frick, Jessica Sutter, James Marsh, to cite merely a few. Also, a big thanks to Karin Klein and Ursula Schach for the administrative and lab support.

Also, thank you to those outside of the department that shared their friendship, student tips, and good vibes: Magdalena Firlej, Isabella Casini, Christian Klask and Christian Fink, Sergio Latorre, Simone Larivera.

Thanks to my previous supervisors who gave me the opportunity to develop my lab skills and scientific culture in their labs: Marie-Anne Félix, Silvia de Monte, Reindert Nijland, and Peter von Dassow.

In general, the Ph.D. road can be bumpy and foggy - but in a car filled with friends, the trip is so much better: Laura, Andrea, Magda, Jorge, Philipp, Armando, Isabella, Nils, Leo, Liam, Sergio, Bilal, Carina, Sara, Marco, Stacey, Gero, Gabriela, Moï, Simone, Nils.

And the ones with whom I was on the road before it even started. Les copains de France: Tiffany poupoune, Tristao, Alex dit Zieder, Jerem, Gui, Johan, Chloé, Pierre'zer, Tancrède, Ahmed, Lucile, Ségo, Simon, MS, Léa, François, Leo, Xavier; the friends from the Netherlands y los weones de Chile. Not in the list? Still in my heart.

And the random ones: las arepas y empanadas (see Borbón et al. for method details), fr**a**nzbrötchen, Damon Albarn, the Beatles, the Clash, la playlist Pas Content, the Tübingen Studenten Philharmonie orchestra, Pint of Science, Mübi, and the Plancktunes.

Et puis surtout, mes parents, ma soeur, mes frères, ma grand-mère qui m'ont soutenue où que je sois. Merci.

General Remarks

In accordance with the standard scientific protocol, the personal pronoun ‘we’ will be used to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

List of Figures

List of Tables

Abbreviations

ALP	Adhesin-Like Proteins
BMI	Body Mass Index
CV	Cross Validation
df	degree of freedom
FDR	False Discovery Rate
FS	Feature Selection
GBM	Gradient Boosted Model
GC	Gas Chromatograph
gRRF	guided Regularized Random Forest
HPC	High Performance Computing
HPLC	High Performance Liquid Chromatography
ICE	Individual Conditional Expectation
MAG	Metagenome Assembled Genome
MB method	Meinshausen and Bühlmann method
ML	Machine Learning
MPI-DB	Max Planck Institute for Developmental Biology
PDP	Partial Dependence Plot
PR	Precision-Recall
RF	Random Forest
RIT	Random Intersection Trees
RMSE, MSE	Root, Mean Squared Error
ROC	Receiver Operating Characteristic
SCFA	Short-Chain Fatty Acid
SEM	Scanning Electron Microscopy
SHAP	SHapley Additive exPlanation
SRB	Sulfate-Reducing Bacteria
TP, TN, FP, FN	T: True, P: Positive, N: Negative, F: False

Prologue

We are made by half of ourselves, i.e., made of our own human cells, and by half of microbes, i.e., made of microorganisms present on our skin and in our gut, for example [204]. Does it matter? Probably more than we imagine. More than 90 % of our microorganisms are localized in our gut [204], at the interface between the undigested food that reaches the intestine and the epithelial cells of the gastrointestinal tract [63]. Their presence alone is important to us as the proteins and glycolipids they harbor on their cell membranes can be recognized by the host, hence mediating immunity [137]. In addition, they degrade undigested dietary components, prolonging our digestion and providing us with additional energy sources. Finally, they produce metabolites that can interact with all kinds of host cells, including T-cells, involved in inflammation; neurons, involved in the gut-brain axis; or adipocytes, involved in energy storage [102].

The gut microbiota has attracted attention in recent years. As the majority the gut microbiome inhabitants are still uncultured, sequence-based studies have the advantage of capturing a broader range of microorganisms without being restricted by the need for isolates [103]. Thanks to advances in sequencing technologies, using stool samples to assess microbial diversity is now cost-effective and allows us to obtain an overall picture of the human gut microbiome diversity. However, most protocols are designed for bacteria, the main microorganisms in the human gut microbiome, therefore overlooking the archaea, fungi and yeasts that also colonize this environment [24, 179].

The major archaeon of the human gut microbiome, in terms of relative abundance and prevalence, is *Methanobrevibacter smithii*, from the *Methanobacteriaceae* family. This family only comprises methanogens, i.e., microorganisms whose metabolism reduces carbon substrates to CH_4 [62]. In the human gut, methanogenesis starts from products of bacterial fermentation; for *M. smithii*, namely H_2 and CO_2, and alternatively formate to the cost of lower energy yield. The H_2 produced during fermentation is also used in other processes such as bacterial sulfate reduction [196]. Hence, methanogens are part of a syntrophic chain starting with host digestion, pursued by bacterial hydrolysis and fermentation, and finally ending with methanogenesis and sulfate reduction [29]. It is hypothesized that by uptaking fermentation products, *M. smithii* participates to modulating the production of bacterial metabolites interacting with the host [46, 196]. Evidence of altered bacterial fermenting activity have been reported for co-cultures of the methanogen with human and ruminant gut bacteria [30, 195]. Furthermore, *M. smithii* has been associated with various host phenotypes using sequencing data [121, 238, 170, 147, 42, 144, 9, 157, 80, 201, 106, 28, 150, 249, 228], likely due to its interactions with fermenting bacteria and revealing insights into its effects on the gut microbiome.

Although published findings shed light on the ecology of methanogens in the human gut, much remains to be unveiled. Methanogens are monitored in cattle to improve food energy intake [82, 223, 111] and, with the emerging idea that they could be used as probiotics for humans [28, 198], it is critical to further our understanding of their role in the human gut. In particular, the ecology of *M. smithii* must be characterized: how it interacts with other members of the human gut microbiome, and what is the ecological feedback between the microbiome and the gut environment.

While sequence-based data are routinely used nowadays in microbiome research [115], the means to properly analyze them are yet to be established. For a long period, classical statistical methods have been employed to infer microbe-microbe and host-microbe associations, although being inappropriate and yielding unacceptable false discovery rates (FDR) [236]. Over the last decade, numerous models have been proposed to fit microbiome data and replace pairwise comparisons with classical statistical analyses [178, 101, 251, 98, 243, 34, 70, 124]. However, these models presuppose that microbial features and their interactions are all of the same kind, and so, can be uniformly modelled by one parametric equation [26]. This is questionable given the complexity of microbial interactions [255], likely explaining the high FDR of microbial network inference techniques based on such models [236, 70].

Accordingly, machine learning (ML) algorithms are emerging as the most appropriate methods to investigate relationships between hosts and microbes [115, 116, 225]. Their flexibility and robustness are adapted to microbiome data and they proved to be accurate for predicting host phenotypes using microbial features [135, 171, 36, 222, 182, 220, 65, 245]. Unfortunately, if their complexity enables capturing complex interactions, it also renders their interpretation more difficult. Methods to apprehend ML models have been suggested but were not designed for fully describing high-dimensional models [140, 190, 123, 47, 14]. Their use with microbiome data is hindered by the high-dimension of sequence data sets and results in limited insights. Accordingly, studies utilizing complex models often only report ranked lists of identified taxa and carry on with their analyses using classical statistical tests [2, 85, 171, 220, 65, 182, 245, 36, 248, 91, 59, 15, 110, 222].

Thus, there is a need for the field to overcome the interpretation issue of complex models. Solutions will allow comprehending interactions between components of the gut microbiome and their host.

Chapter 1
Background

1.1 The human gut methanogen *Methanobrevibacter smithii*: description and detection in human guts

1.1.1 *M. smithii* depends on fermenting bacteria to colonize the human gut

Methanobacteriaceae are the most prevalent and abundant archaea in the human gut [23, 97]. Together with *Methanomassiliicoccales*, another clade of archaea detected at considerably lower abundances and prevalence, they compose the pool of methanogens found in the human gut [24]. Methanogens produce CH_4 as an end-product of their metabolism, through the reduction of CO_2, methylated compounds or acetate [62, 156]. H_2 is generally the electron donor for the reaction, and can be sourced from the environment or generated during steps prior to the reductive step [62, 156, 134]. While *Methanobacteriaceae* are CO_2-reducing archaea, beside the genus *Methanosphaera* that can also start methanogenesis from methanol [62, 156, 134], *Methanomassiliicoccales* require methylated substrates [21, 22, 56]. Furthermore, certain methanogens can cleave formate into CO_2 and H_2, allowing them to use formate as an alternative to H_2 and CO_2 [62, 156, 134, 94].

Methanobrevibacter smithii accounts for the great majority of

Methanobacteriaceae in the human gut, both in terms of prevalence and abundance [23]. It has been associated with various host phenotypes such as constipation and slow transit [121, 238] or non-western diet [170, 147, 42]. Interestingly, relative abundances of *M. smithii* have also been correlated with low [144, 9, 157, 80, 201, 106, 28] and high body mass indexes (BMI) [150, 249, 228]. If these contrasting observations are correct, then the context in which each process occurs must differ. For instance, methanogens may promote a high BMI in the context of one particular diet or microbiome composition, while promoting leanness in the context of another. This hypothesis is supported by the dependency of *M. smithii* upon bacterial fermentation for its methanogenesis. Non-digested carbon sources that reach the colon are degraded through fermentation into short-chain fatty acids (SCFA), H_2 and CO_2 by gut bacteria (Figure 1.1). Produced SCFAs mostly comprise acetate, propionate, and butyrate, which can be absorbed by the host, where they mediate energy metabolism by means largely described in [29, 161, 138, 46, 120, 137]. When enzymatic activity is not limited (e.g., low enzyme or cofactor concentration), the second law of thermodynamics predicts that a chemical reaction is boosted by removing products [118]. Therefore, by uptaking H_2 and formate, *M. smithii* may promote specific bacterial fermentation pathways and alter SCFA production, subsequently mediating host metabolism (Figure 1.1). For instance, in co-cultures with the H_2-producer *Ruminococcus flavefaciens* or with the formate producer *Fibrobacter succinogenes*, the methanogen promotes acetate production [195]. However, in cocultures with *Ruminococcus albus* or *Roseburia intestinalis*, its H_2 consumption does not trigger any change in SCFA production and therefore in bacterial fermentation activity [195, 30].

Several fermentation pathways occur in the gut, all of which involve disparate substrates, products, and microorganisms (Figure 1.1) [138,

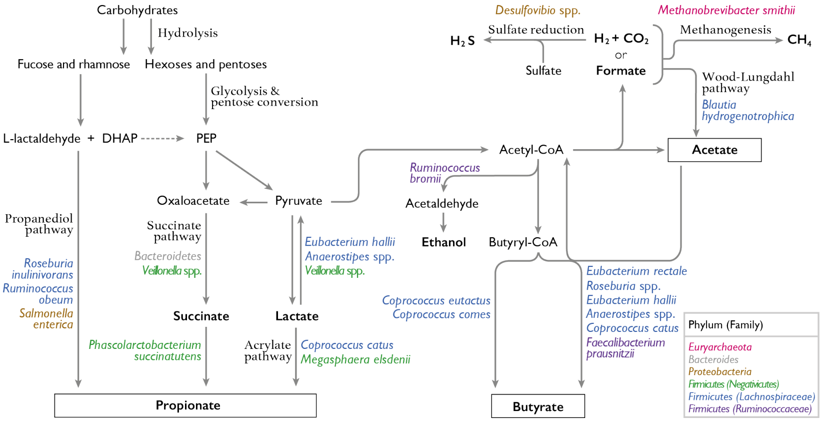

Figure 1.1: Metabolic pathways performed by microorganisms in the human gut, from carbohydrate degradation to CH_4 production. Names of metabolic pathways are indicated in a distinct font and main metabolites are indicated in bold. Microorganisms are colored by taxonomic phylum or family; taxa shown are not exhaustive and based on literature. DHAP: dihydroxyacetonephospate; PEP: phosphoenolpyruvate. Figure adapted from Louis et al. [138].

46]. The dependence of *M. smithii* on H_2 and formate may drive its adaptation to the human gut microbiota [94, 198]. *M. smithii* has acquired the formate dehydrogenase (Fdh) via horizontal gene transfer [94], which allows the archaeon to start methanogenesis from formate. Although the use of formate is energetically less efficient [134], it provides an alternative to escape competition with hydrogenotrophs, i.e., sulfate-reducing bacteria (SRB) and acetogens, when resources of H_2 are scarce (Figure 1.1) [196, 213, 168]. *M. smithii* strains also harbor a great variety of adhesin-like protein (ALP) genes coding for proteins thought to mediate direct cell-cell interactions [94]. Fick's law of diffusion dictates that the efficiency of interspecies H_2-transfer substantially increases with decreasing distance among cells [208, 212]. This diversity in ALPs may be advantageous to *M. smithii* to establish direct contact with H_2-producers, as observed for *Methanobrevibacter ruminantium*, a methanogen closely related to *M. smithii* and colonizing ruminant guts [128, 165].

As the ability of *M. smithii* to colonize the human gut partially depends on the gut microbial composition and metabolism, its prevalence and relative abundance in human populations must be associated with members of the human gut microbiota, metabolic functions, and host habits likely to alter the host microbiome [193, 234]. Some associations have already been reported. For instance, *M. smithii* co-occurs with specific H_2-producers, such as members of the *Christensenellaceae* family [80, 94, 230, 114] and cellulose-degrading *Ruminococcus* spp. [31]. Carbohydrate intake also positively correlates with a microbial community composed of fermenting bacteria, fungi, and the *Methanobrevibacter* genus [97], and the addition of resistant starch to diet is associated with an increase in fermenting bacteria and *M. smithii* [230]. More generally, the methanogen is associated with specific microbiome profiles as defined by the enterotype landscape. While it is depleted in microbiomes enriched in the *Bacteroides* genus, it is more abundant in microbiomes with the highest *Prevotella/Bacteroides* relative abundances ratio [10, 38, 97].

1.1.2 *M. smithii* is a fastidious organism, limiting its use in culture-based approaches

M. smithii has a slow replication time: as I show in Chapter 5, according to measurements of gases under optimal growth conditions in a rich medium with excess methanogenesis substrate, *M. smithii* reaches stationary phase after five days of growth, while the commensal gut bacteria *Bacteroides thetaiotaomicron* and *Christensenella minuta* reach this phase in less than two days (Figure 5.1). As a result, culture experiments of the methanogen with other bacteria in batch mode, i.e., without continuous renewal of growth medium, may be suboptimal due to the asynchronous growth of the micro-organisms involved. Continuous experiments necessitate bioreactors, which are expensive and time-consuming equipment. Therefore, they often do not allow multiple replicates and constrain the experimental design to a minimal number of tested conditions [226, 60]. This means that researchers face a trade-off between more biologically relevant continuous experiments with very few replicates and more comprehensive batch experiments with several tested conditions, controls, and replicates.

Methanobacteriaceae are obligate anaerobes, i.e., they are sensitive to O_2 [12]. For this reason, specific culture conditions are required to manipulate them. The growth medium must be boiled or sparged with N_2 to remove O_2 and kept under an O_2-free atmosphere. Likewise, inoculated cultures must be handled under an O_2-free atmosphere and with O_2-free material, the most optimal being the use of anaerobic glove-boxes [12]. Furthermore, *M. smithii* requires an external source of H_2 for growth [155]. H_2 is a flammable gas and so it cannot be contained in high concentrations and high volumes in laboratories. Due to the need for a growth atmosphere that is both O_2-free but H_2-containing, cultures are commonly grown in Balch tubes pressurized to 1 or 2 bar with a 80/20 % v/v H_2/CO_2 atmosphere [12]; though recently, protocols including bacteria as an H_2-source have been proposed [113, 227]. High-throughput screening machines, e.g., plate readers, are not designed to be compatible with such tubes. In addition, equipment designed for high-throughput screening exposes microorganisms to concentrations of H_2 too low for *M. smithii* to grow in optimal conditions [146]. Consequently, fastidious methods necessary to grow methanogens limit their inclusion in culture-based investigations of the human gut microbiome.

1.1.3 Shotgun metagenome sequencing is the most appropriate method to infer archaeal ecological patterns from human gut microbiome studies

The prevalence, abundance, and diversity of archaea in human gastrointestinal tracts are commonly underestimated [24]. The first historic method utilized to determine the prevalence and abundance of archaea in human guts was the quantification of CH_4 concentrations in host breadth. However, the majority of the CH_4 produced in the human gut is excreted through flatulence [156], leading to a mis-calculation of methanogen's prevalence and abundance.

The second method consists of estimating the relative abundance of methanogens in stool samples from sequence data. Although more reliable than breath CH_4 concentrations [61], such methods still involve certain limitations [179, 24]. Archaeal cell walls are composed of pseudo-peptidoglycans that lyzozyme fails to disrupt [62, 13]. Since, DNA extraction protocols often rely on lyzozyme to free DNA from cells, they may not be adapted to archaeal physiology and samples may necessitate additional treatments to recover as much archaeal DNA as possible [54, 197, 179]. Furthermore, the choice of primers is crucial for 16S rRNA gene amplicon studies [179]. For instance, primers targeting the V1-V2 regions of the 16S rRNA gene do not hybridize with archaeal DNA, resulting in a failure to detect archaea in taxonomic profiles generated with such primers [122]. Finally, although primer limitations do not apply to shotgun metagenome sequencing procedures, archaea's general lower abundance compared to bacteria [179, 226] and their under-representation in genome databases [211] can also result in an under-estimation of the archaeome importance and diversity in the human gut [24].

In conclusion, analyses of shotgun metagenomes are currently the most appropriate method to explore associations between human gut methanogens with host factors and co-occurring bacteria. Thanks to efforts to enrich genome databases [43], they have already extended the known archaeal human gut diversity [23, 177, 44].

1.2 Analysis of shotgun metagenome sequences from gut microbiomes

1.2.1 Assessing taxonomic diversity from metagenomes

Metagenomes constitute the ensemble of DNA isolated from an environment [92]. When applied to the human gut microbiome, shotgun metagenome sequencing consists of extracting all DNA from stool samples, sequencing sheared DNA fragments and processing the obtained sequences [77, 235, 186]. Sequences are first filtered based on quality scores, e.g., length and coverage, and human sequences are removed [108]. Once cleaned, the remaining sequences can be mapped onto reference genome databases to infer taxonomic microbial abundance [104, 239, 105, 203, 139, 43] or metabolic profiles [1, 67]. Sequences can also be assembled and binned, e.g., according to their relative abundances or sequence similarity, to establish gene catalogs [167, 184] or build metagenomic assembled genomes (MAG) [210, 112, 242]. Bins can be mapped to genome databases to identify the taxa to which they belong and obtain taxonomic abundances [186, 73]. MAGs assembly is beneficial to uncover diversity [174, 177, 162, 6, 66, 252] and allows genome databases to be extended to include uncultivated microorganisms [43, 32, 7].

The aim of investigating metagenomes is often to associate variables, taxonomic or metabolic profiles, to a phenotype. However, metagenomes have intrinsic properties that complicate their analyses. In the following sections, I first review those characteristics and then evaluate current analysis procedures.

1.2.2 Considerations for the analysis of metagenomes

Hierarchical features Microorganisms are named according to a taxonomy that follows their phylogenetic tree. Limits that divide microorganisms into taxonomic groups are arbitrarily defined and do not consistently reflect metabolic capacities or specificities [206, 172]. Consequently, describing microbial diversity with a unique taxonomic level may not capture microbial interactions in a community. For example, let us assume a metabolic function is shared by members of the same taxonomic family. A phenotype dependent on this function may thus be strongly correlated with the family relative abundance. However, at the genus level, the correlation may be weaker due to the ability of every genus to perform the function, hence attenuating the statistical signal. When investigating microbial interactions, choosing an inadequate taxonomic rank can lead to biased results. Finer ranks will provide a better resolution that will distinguish specialized interactions, but they may miss generalized interactions. Conversely, coarser ranks will allow to identify general patterns but not specific interactions. The same issue applies to genes, metabolic pathways, and other objects described by metagenomes.

High-dimensionality The dimensionality of a dataset corresponds to the number of variables p describing the set of n observations; a dataset is high-dimensional when $p >> n$ [96]. Each observation can be seen as a vector of values of the p variables. This vector is one among all possible vectors from a space of p-dimensions. As p increases, the set of n observations will represent a relatively smaller set of the p-dimensions space. It will thus be harder for models to evaluate the general association of each variable or even interactions of variables with the phenotype, and associations detected may be true only for the specific set of samples used for analyses. Overfitting means producing results that are not generalizable. The risk of overfitting analyses to the set of observations is amplified when p increases, which is called the *curse of dimensionality* [96]. Although certain statistical methods can help to analyze high-dimensional data, reducing p before analyses remains necessary to obtain the best results [96]. The choice of taxonomic ranks to include in analyses is therefore dependent on the number of samples to keep $p < n$, or at least as close to n as possible. As microbes are organized into functional networks [254, 160, 138, 137], their interdependencies may cause a combination of taxa to be correlated with the response variable, but not each taxon of the group. Therefore, interactions should be included when investigating microbiomes at the cost of increasing dimensionality.

Inter-dependence and redundancy A drawback when considering including several taxonomic ranks in analyses is the redundancy in the information of nested levels. Relative abundances of taxa from the same phylogenetic branch may be highly correlated as they represent subsets of the same microorganisms [172]. Variables carrying very similar information would thus not only increase dimensionality and noise, but also subsequently reduce the power to distinguish important features.

Compositionality The abundance calculated from metagenomes is bounded by the sequencing depth of the extracted DNA. Consequently, it is the relative abundances of taxa that are computed. Metagenomic data are therefore compositional: if the absolute abundance of a taxon increases, the increase in its relative abundance is automatically associated with a decrease in the abundances of all other taxa [78].

Sparcity Human gut microbiomes are highly diverse and vary with age, geography, and many other factors [177, 193, 10, 117]. Accordingly, microorganisms are not ubiquitous and are often present at low prevalence and abundances. When comparing large heterogeneous human populations, this microbial disparity generates zero-inflated variables with distributions skewed towards zero and different variances.

Human factors In addition to the aforementioned technical biases, e.g., DNA extraction protocols, sequencing depth, or reference databases for mapping reads, biases in studies may arise from sampled populations. Since the environment has a significant influence on the human gut microbiome [177, 193, 117], host co-factors, e.g., alcohol consumption, age, and diet restrictions, can confound microbial relative abundances or the studied phenotype, leading to a bias in results [234].

1.2.3 Methods for comparing metagenomes

In this section I present common methods used to identify microbial taxa and functions associated with phenotypes, e.g., diseased versus healthy host.

Pairwise comparisons Classical pairwise comparisons with p-value correction are widespread in microbiome studies. Correlation tests examine whether two numeric vectors have a linear correlation, e.g., Spearman's coefficient of correlation, if numeric values of one vector are relatively higher for samples belonging to a grouping factor compare to another group, e.g., Wilcoxon-rank-sum tests, or if the distribution of groups across factor variables is not as expected by random, e.g., χ^2-tests. Subsequent p-value correction methods, such as the Benjamini-Hochberg method [16], aim at accounting for the increased chance of obtaining a significant correlation when testing multiple associations and variables. However, results from these studies are open to criticism as such statistical tests are not designed for compositional zero-inflated data and cofactors cannot be taken into account [115]. In support to these points, the parametric Pearson and non-parametric Spearman tests, followed by Bonferroni correction, have been shown to yield an unacceptably high false discovery rate (FDR) [236]. Furthermore, pairwise comparisons cannot evaluate variable interactions associated with the phenotype of interest. These methods are thus not recommended for exploring metagenome profiles.

Generalized linear models Generalized linear models predict a variable of interest, also named response variable or target, by approximating its distribution with other variables, also named predictors or features. The fitted distribution of the response variable can be adjusted to better correspond to metagenome data. For instance, models have been developed to account for sparcity by using distributions skewed towards zero such as the zero-inflated Gaussian [178], beta-binomial [101, 251], or Dirichlet-multinomial [98, 243] distributions, to cite merely a few. An advantage of linear models is that cofactors due to experimental design, e.g., batch effects, or sampled population, e.g., age, can be included in the set of predictors to correct for their effect. Mixed effects models have been proposed to better take into account cofactors [87, 34]. In such models, cofactors with an expected consistent effect such as drug usage are included within the set of predictors as fixed variables, while others are included as random variables that are assumed to describe correlations between observations [151]. For instance, in a longitudinal study, sampled individuals would be included as random effects to account for the dependency of their samples [34]. Furthermore, the simple structure of linear and generalized models facilitates their interpretation. One parameter is fitted for each predictor, or pair of predictors, such that their effects on the response variable can easily be visualized. However, due to their simplicity, such models are limited for capturing complex feature interactions. Their rational is also based on the idea that data follow the fitted model, hence assuming that data were generated in a relatively mechanistic manner [26].

Covariance matrices Sparse covariance matrices are used in microbiome science to determine conditionally non-independent taxa and build correlation networks [69]. The comparison of networks computed from distinct sample groups can help identifying associations specific to these groups. For instance, by comparing networks generated for samples collected from environment A versus environment B, one can infer associations of variables specific to each habitat [70]. Several methods exist to estimate covariance matrices, all proposing different approaches to deal with the compositionality and sparsity of metagenome data.

- The sparCC algorithm calculates the covariance of variables across observations. Relative abundances are first adjusted with a pseudocount to avoid zero values, are subsequently log-transformed to reduce the effect of compositionality [3], and the covariance of pairs of variables is finally estimated [70].
- Meinshausen and Bühlmann, 2006 [152] proposed an algorithm which first fits LASSO linear models to each pair of variables and then uses the estimated parameters to create the covariance matrix. The lasso regularization consists of including a penalization parameter β_i for each variable i in the model to decrease variable effects. The β_i parameters will be null or very low for most variables, such that only a few variables will be used in the final model, therefore forcing sparsity [224]. In the Meinshausen and Bühlmann (MB) method, models are fitted for each variable given the other variable in the pair. For each pair, if one or both lasso parameters are not null, variables are assumed to be correlated [152].

- The graphical lasso method [69] builds upon the MB method, which is suggested to be a simpler approximation of the covariance matrix [69]. Hence, similarly to the MB method, graphical lasso fits lasso models and uses the estimated penalization parameters to make the covariance matrix. However, variables are assumed to follow a Gaussian distribution and the fitting procedure is iterated to find parameters that maximize the log-likelihood of all variables to follow such distribution.
- Both the MB [152] and graphical lasso [69] methods are implemented in the SPIEC-EASI framework [124]. To account for compositionality, a centered-log ratio transformation is first applied to relative abundances of taxa [3, 124]. Covariance matrices are then computed with either the MB or graphical lasso method.

A drawback of covariance matrices is their incompatibility with categorical variables. Therefore cofactors cannot be included in analyses, hence limiting the application of covariance matrices to studies where experimental design or population structure may exist.

Tree ensemble models Tree ensemble models are a class of machine learning (ML) models that rely on decision trees to predict a response variable [96]. As for linear models, the response variable can be numeric, e.g., relative abundance of a taxon, or categorical, e.g., presence/absence of a taxon. Predictors can also be of any sort, i.e., numeric or categorical. Decision trees are structured to split sets of samples into more homogeneous subsets in terms of response variable. Diverse procedures have been proposed to build trees, with the random forest (RF) algorithm being the most used in microbiome science [25, 115, 116]. General considerations, unspecific to metagenomes, are detailed in the next section.

The sequential splitting of samples by different predictive variables enables tree ensemble models to capture complex interactions between predictors. In addition, contrary to linear or generalized models, no assumption is made on data distribution with tree ensembles. RFs, a class of tree ensemble models for which trees are grown independently from each other [25], have also been shown to be robust to high-dimensional data [116, 115, 225, 250]. Therefore, tree ensembles are compatible with sequence data and are the most appropriate models to analyze metagenomes, to date. However, their use in microbiome science is still limited due to their complexity and consequently low interpretability. Although they achieve good predictive performances with host phenotypes using relative abundances of microorganisms [116, 115, 225, 250], the underlying variable interactions are commonly summarized by a measure of the variable importance in the model, resulting in only a list of taxa important for predictions.

1.3 Fitting of tree ensemble models for predicting host traits with metagenomes

A *model* can designate the predictive object, e.g., an RF fitted on a given set of observations and predictors, or the processes that produce such object, e.g., the algorithm and set of parameters utilized to obtain the predictive object. To fit a good model, one must (i) select the best sequence of processes for fitting the final predictive model, and (ii) measure the quality of the final model, i.e., its *generalization* to new data. As rightfully stated by Pedro Domingos [51], it is not fitting the final model that takes time but selecting the model to fit. In the following paragraphs, I will briefly describe the pivotal points to consider for fitting tree ensemble models to microbiome data. My main resource to write these paragraphs was the book “The elements of statistical learning: data mining, inference, and prediction” from Hastie, Tibshirani, and Friedman (2009) [96] that exhaustively present all considerations and procedures for fitting ML models. It was notably written by three of the researchers whose work provided a foundation for the field.

1.3.1 Model assessment

The ability of models to make accurate predictions depends on several factors. First, many ML algorithms exist: RF, gradient boosted models (GBM), linear regressions, and neural networks, to cite only a few. Each model has a different structure and learning process that will affect its ability to predict the analyzed data. In addition, steps upstream to model fitting, e.g., feature selection, modify data and consequently the predictive performance of models. Furthermore, models rely on hyper-parameters which value often control the model complexity, e.g., the number of trees in a RF, or the learning process, e.g., the loss function used in boosted trees. The risk with increasing model complexity is to overfit models to the input data. Overfitting means that predictions are accurate only for observations very similar to the set of data used to fit the model. Consequently, an overfitted model does not perform well on new observations.

Models are assessed by measuring their predictive accuracy on unseen observations, which were not used to fit the model. Hence, a validation set of samples is kept aside when fitting models and is predicted afterwards. The model with the highest accuracy on the validation set is selected to fit the final model. Ideally, the measure of the quality of the final model would be performed on an independent unseen data set. However, when the number of observations available for analysis is limited, cross-validation (CV) is instead routinely employed for both selecting a model and estimating its expected accuracy. For this, model training and testing is repeated on different splits, or folds, of the full data set. In each fold:

1. observations are assigned to either the train or test set,
2. a full model , i.e., the complete sequence of processing steps for a given ML algorithm with fixed hyper-parameter values, is fitted on the train set,
3. the model accuracy is measured on the test set.

The accuracy of each full model is averaged across folds to estimate the generalization of the model and guide model selection. The selected model is finally fitted on all observations.

In practice, five or ten folds are recommended for CV, and the sample sizes of train-test sets depend on the noise in data and total number of observations available [96].

1.3.2 Model quality

The quality of a model is estimated from its predictions on new data for which the truth is known. The function utilized to measure the quality depends on the goal of the model and data used to fit the model.

Let $y \in \mathbb{R}$ represent a response variable observed on n samples, such as y is continuous for regressions and 0-1 encoded for classification. Classifications of multi-class variables can be simplified into binary problems by choosing a positive class encoded by 1 and assigning 0 to all other classes. Let $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ be the average response variable across all observations and $\hat{y}_i$ be the prediction of the model on a sample $i \in \{1, \dots, n\}$.

Regression

For regression, the goodness of predictions is typically measured using the deviation of predictions from the true values [95, 96, 214], e.g., with the root mean squared error (RMSE) or coefficient of determination

R^2 [53],

$$\text{RMSE} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{(y_i - \hat{y}_i)^2} \quad \text{and} \quad R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\ .$$

The smaller the average deviation, the more accurate the predictions.

For certain tasks, more sophisticated functions may be more appropriate to measure the quality of models. For instance, *survival* analyses investigate the occurrence of events in time, e.g., looking at the remission of patients according to treatment. Thus, models are fitted to predict the risk or probability of the event for each observation of each individual, and the concordance statistic (c-statistic) can be used to assess the quality of models [214, 95]. This index compares for pairs of individuals, the probabilities predicted for individuals who experienced the event versus those predicted for individuals who did not. For instance, the probabilities of remission for patients that received a placebo should be lower than for patients who received the treatment. The higher the c-statistic, the better the model is at discriminating individuals at risk [214].

Classification

For classification, measures are based on a confusion matrix which counts the number of well-classified observations within each class, i.e., the true positive (TP) for the 1 encoded class and true negative (TN) for the 0 encoded class (illustrated in Table 1.1).

	Positive prediction	Negative prediction
Observed positive class	True positive (TP)	False positive (FP)
Observed negative class	False negative (FN)	True negative (TN)

Table 1.1: Confusion matrix for binary classification

Numerous metrics are derived from the confusion matrix (Table 1.1), each describing different proportions of well- or mis-classified samples, and being used alone or in combinations to measure the quality of classifiers.

First, the *accuracy* is commonly used to assess the overall predictive performance of a model by looking at the ratio of well-classified samples over all samples, Equation (1.1). An accuracy between 0.5 and 1 means that a majority of samples were well-classified, and so that the classifier does better than random. However, this metric is not fair when classes are imbalanced. If only the majority class has been well learnt by the model, most samples in this class would be well-classified, hence resulting in a large accuracy despite the minority class being poorly predicted.

$$\mathcal{A} = \frac{\mathrm{TP} + \mathrm{TN}}{n} \tag{1.1}$$

Consequently, Cohen's κ [37] is recommended to evaluate the quality of a model for imbalanced data. This metric compares the proportion of well-predicted samples of each class to the predictions that would have been expected by random, i.e., by a random guessing classifier. A model predicts a ratio $\frac{\mathrm{TP}+\mathrm{FP}}{n}$ of samples to be in the positive class; while by random, the proportion of observed samples in the positive class $\frac{\mathrm{TP}+\mathrm{FN}}{n}$ is expected. Thus, the probability for a sample to be well-classified in the positive class by the model but also by a random classifier is $p_{y=1} = \frac{\mathrm{TP}+\mathrm{FP}}{n}\frac{\mathrm{TP}+\mathrm{FN}}{n}$. Likewise, the probability for a sample to be well-classified in the negative class by the model and random classifier is $p_{y=0} = \frac{\mathrm{TN}+\mathrm{FN}}{n}\frac{\mathrm{TN}+\mathrm{FP}}{n}$. Cohen's κ is defined as

$$\kappa = \frac{\mathcal{A} - (p_{y=1} + p_{y=0})}{1 - (p_{y=1} + p_{y=0})}\ ,\ \kappa \in [-1, 1]. \tag{1.2}$$

A classifier is better than random guessing when $\kappa > 0$ and, as a rule of thumb, $\kappa \geq 0.6$ indicates a good classifier [126].

The receiver operating characteristic (ROC) curve is an alternative to Cohen's κ which is particularly useful to tune the model classification threshold. Similar to the ROC curve, the precision-recall (PR) curve focuses on the well-prediction of one particular class instead of both. These curves are built by varying the threshold used by a model to predict a class. For instance, RFs calculate an average score from all predictions made by individual decision trees. A threshold is then employed to discriminate classes and assign a label to samples according to their score. Consequently, if we consider a model fitted to detect cancer from patient material, a sample with $\hat{y}_i = 0.8$ may be classified as healthy for a threshold of 0.5, but as cancer for a threshold of 0.9. The ROC curve shows for each threshold the *recall* and *specificity*, Equations (1.3) and (1.4), respectively. The recall, also named *sensitivity*, measures the proportion of well-classified samples from the positive class, i.e., the true positive rate. The specificity mirrors the recall for the negative class. As the name suggests, the PR curve shows the *precision* and recall, Equations (1.5) and (1.3), respectively. The *precision* measures the proportion of TP among samples classified as positive by the model; it corresponds to 1 - the false discovery rate (FDR), Equation (1.6).

$$\mathcal{S}n = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (1.3) \qquad \mathcal{S}p = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \qquad (1.4)$$

$$\mathcal{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (1.5) \qquad \mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} = 1 - \mathcal{P} \qquad (1.6)$$

Metrics aforementioned are the most common ones found in the literature, though many others exist. The choice of the model quality measure depends on the set of observations used for model fitting, on the task performed, and on the purpose of the model.

1.3.3 Tree ensemble algorithms

Tree ensemble models make predictions from decision trees: predictions made by each tree are gathered to make the final prediction. Different strategies can be employed to grow the ensemble of trees, RF and GBM are the two main methods that I will detail in the following paragraphs [96]. They come in various variations and are the building blocks of tree ensemble ML models [25, 71].

Decision trees

A decision tree splits a set of samples into more homogeneous subsets in terms of response variable [96]. The mean squared error (MSE) and Gini impurity are the most common error criteria used to measure the homogeneity of subsets for regression and classification, respectively [25, 71]. The Gini impurity, for generalized inequality index, sums the probabilities p_k of randomly drawing each class $k \in K$ of the response variable from a set of observation [218],

$$\text{Gini} = -\sum_{k \in K} p_k(1 - p_k) \ . \tag{1.7}$$

For a binary 0-1 encoded response, this becomes Gini $= 2 \cdot p_{y=1}(1 - p_{y=1})$, such as Gini $= 0$ when a unique class is found in the set of observations, and Gini $= 0.5$ when both classes are found in same proportions. The MSE compares response values to the average across the n samples $\bar{y}$,

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \ . \tag{1.8}$$

The improvement due to a split on n observations is calculated such as the error criterion measured before splitting is compared to the sum of error criteria measured on the two created subsets R_l and R_r, weighted by their respective subset sizes $|R_l|/n$ and $|R_r|/n$.

A decision path is then defined as an ordered sequence of splits (a branch of the tree); decision paths partition observations into non-overlapping subsets from which predictions are estimated by taking the average response value. The number of splits in a decision path corresponds to the interaction order, such as a decision path consisting of 1 split corresponds to a main effect.

Random forests

RF algorithms grow independent decision trees on bootstraps of sample sets. Each tree makes a *vote* on predicted observations, and votes are averaged across the forest to obtain the final predictions [25].

Within trees, at each node a set of $m \leq p$ predictive variables is randomly drawn. The algorithm then searches for the predictor and threshold value that separate samples into two subsets with the minimal response variable variance. The procedure is repeated for each newly created subset. The variance can be measured by different functions, the MSE and Gini impurity being commonly used for numeric and categorical responses, respectively [96, 25].

Consequently, when $m = p$, similar trees may always be grown and the deviation between predictions and truth may be high due to a smaller exploration of the p dimensions. Conversely, if only a very small number of predictors is drawn at each split, irrelevant predictors may be used and therefore given importance. The rule of thumb is to use $m = \lfloor \sqrt{p} \rfloor$ for classification and $m = \lfloor p/3 \rfloor$ for regression [96].

Fully grown trees have tips consisting of homogeneous sets of observations, i.e., with a null or very low variance. Models with fully grown trees or comprising a high number of trees are more complex, which can lead to overfitting. However, as trees are grown independently on sample bootstraps, RFs are robust to tree depth and forest size, such as not much tuning is needed to achieve a good model [96, 25, 26].

Nonetheless, a simplicity bias can be introduced by pruning trees or limiting the size of the forest with a minimal loss of accuracy [96, 202].

Gradient boosted models

While in RFs decision trees are grown independently from each other on bootstraps, on the contrary, the boosted method learns non-independent trees on all data. The algorithm learns to predict data by adjusting itself in an iterative fashion while minimizing a loss function [71, 96]. The loss function measures the deviation of the predictions to the real observed values. The idea of boosting is to fit weak models to predict samples that were not well-predicted by models from previous iterations and add them to obtain a strong final model. In comparison, RF aggregates perfect models fitted on bootstraps to obtain a strong average model.

In essence, GBMs are constructed as follow. Let $m \in \{1, \ldots, M\}$ be the iteration number, T_m be the decision tree fitted at iteration m, and f_M be the GBM. The algorithm starts by fitting a decision tree T_1 on all samples. *Gradients*, or *pseudo-residuals*, derived from the loss function are calculated, and a new tree T_2 is fitted on these residuals. Each decision path $R_{j,2}$ of T_2, $j \in \{1, \ldots, J\}$ and J the number of leaves in the tree, is then weighted so that to minimize the loss function of the model comprising both T_1 and T_2. The procedure is repeated for all M trees. The final model corresponds to

$$f_M(\mathrm{x}) = \sum_{m=1}^{M} \gamma_m T_m(\mathrm{x}) = \sum_{m=1}^{M} \sum_{j=1}^{J} \gamma_{jm} \mathbb{1}_{R_{j,m}}(\mathrm{x}) \ ,$$

with $\gamma_{jm} \in \gamma_m$ the weight for each decision path $R_{j,m}$ in T_m. Therefore, a GBM is the addition of several weighted decision paths, fitted sequentially by focusing at each iteration on samples not yet learned, i.e., with absolute gradients larger than zero. The complexity of GBM increases with the number of trees (or number of rounds) and tree depth. Here, since the model is additive, the variance increases with the variance due to each tree and so with complexity. In practice, the tree depth J is advised to be $4 \leq J \leq 8$ and the number of round should be increased until no further gain in accuracy can be achieved [96].

1.3.4 Optimizing models with feature selection

As we have seen in the previous section, tree ensembles are built by testing the ability of predictive variables to partition the response variable. Intuitively, as p increases with irrelevant variables, the probability for the algorithm to select only relevant variables decreases. Consequently, the resulting model will be more complex due to the high number of variables utilized, and more noisy due to the use of variables randomly associated with the response for a few observations. Feature engineering is hence important for model fitting and can greatly improve model quality [116]. However, it is the most difficult part since it is mostly knowledge based.

First, variables can be filtered in a reasonable manner by removing all that are most likely to be noisy. For instance in microbiome studies, relative abundances of rare taxa, i.e., with low prevalence and average abundance, may be associated with the response variable only by random due to an under-representation. Hence, such taxa are routinely excluded from analysis.

Further filtering can be performed by estimating the association of predictors with the response variable and many algorithms have been proposed to this end [116, 45]. Pairwise tests such as Spearman or χ^2 tests can be used to remove variables. However, such tests do not account for variable interactions and may thus exclude important features that do not have a main effect [51]. Consequently, complex models such as RFs may be used to select important variables before fitting the model [90, 123, 48, 100, 107, 8, 172]. For instance, the Boruta framework fits numerous RFs to compare the Gini importance of features to the the Gini importance of randomized variables [123]. Variables with an average importance higher than randomized features are selected by the algorithm, such that all important features, event redundant ones, will be included. Conversely, the guided Regularized Random Forest (gRRF) algorithm uses a two-step approach to select only the relevant non-redundant features [48]. For this, a RF is fitted and variable weights are computed from the Gini importance. A second RF is then built in a regularized manner, such that the decrease in Gini impurity due a variable, Equation (1.7), is weighted to alter the probability of non-relevant variables to be selected [100] (see also Chapter 4, Equation (4.1)).

Finally, hierarchical features such as taxa can be filtered by taking into account structures in variables. In their Hierarchical Feature Engineering (HFE) framework, Oudah and Henschel [172] propose to first reduce the dimensionality by filtering correlated parent-child features, and then select important features. Hence, in a first step, the Pearson coefficient of correlation is calculated for each pair of parent-child variables, and if it is greater than a user-defined threshold, the child variable is removed. In a second step, a decision tree is constructed to entirely reflect the hierarchical variable structure, e.g., the taxonomic tree. The decrease in splitting error is measured for each node and compared to the averaged splitting error of all parents nodes, e.g., the average across coarser taxonomic ranks. If it is lower, then the child node is considered as irrelevant given the importance of its coarser levels and is removed [172]. While this procedure is elegant, it only aims at reducing hierarchical redundancy from features and does not accept variables from outside the hierarchy, e.g., host metadata. Furthermore, it does not integrate variable interactions and, similar to pairwise tests, may thus exclude taxa that are important only in combination with others. Since microorganisms interact in many processes [254, 160, 138, 137], it is reasonable to assume that such inter-dependent associations with the response variable may exist in microbiome data.

1.3.5 Adapting models to imbalanced data

In medicine, it often happens that cohorts cannot be fully controlled, hence resulting in more healthy individuals than diseased ones for example. In that case, if the aim of the study is to generate diagnosis tools from the cohort, it is important to take into account this imbalance between healthy and diseased cases to make sure that next individuals will be well diagnosed and taken care of. If not, a predictive model trained on imbalanced data will likely be biased towards the over-represented class, also called the majority class. Due to fewer observations belonging to the under-represented one, the minority class, fewer classification rules can be learned [216, 136]. As a consequence, the model would typically predict new observations to be of the majority class, resulting in higher mis-classification rates for samples of the minority one. With the diseased versus healthy example, this would lead to patients classified as healthy although they are not.

Three approaches have been proposed to tackle this issue [136]. First, data can be directly harmonized by down- or over-sampling, i.e., randomly removing samples from the majority class or duplicating samples from the minority one [136, 33, 216]. Second, the *cost-sensitive* method corresponds to adapting model algorithms so that the learning process is tuned to well predict both classes [33, 180, 50, 68, 136]. In this method, a weight is given to classes and the cost of mis-classifying one or the other class is then proportional to the provided weights. Finally, model performance can be evaluated using measures reflecting mis-classification of imbalanced data, such as Cohen's κ, Equation (1.2) [136, 216, 37, 57]. While the second technique requires using specific learners, the two others are compatible with any method. Furthermore, the first and second options intervene in the learning process and are thus useful to improve models. Evaluation measures reflecting mis-classification imbalance enable selecting the most appropriate model.

In particular for RF models, the ranger R-package [240] implements options to perform weighted bootstrapping, and so fit decision trees on balanced data, or to attribute weights to classes to adjust the learning process to better learn the minority class. In GBM, gradients of observations can be weighted to reflect data imbalance and inflate their weight during the fitting of decision trees.

1.3.6 Model interpretation

At this stage, models including different feature engineering processes, algorithms and hyper-parameters should have been trained and tested, and compared with an appropriate metric. The final model should have been selected and applied to all data. The expected quality of the model is the one measured during CV.

A few methods can help deciphering models according to needs and model purposes. When the goal of a model is to make accurate predictions on new observations, e.g., for diagnosis purposes, the interpretation may be rather simple: features used by the model are sought to adapt future data collection. However, if the goal of the model is to capture complex interactions from data and describe them with regards to the predicted variable, e.g., to understand why a treatment succeeded with certain but not all patients, more insights into the model are needed than only the list of utilized features. Despite the differences in method bases, the aim is invariant: to provide an estimation of the magnitude of the use of features by the model for predictions. A large feature importance means that predictions heavily rely on this particular feature, and conversely, features with low feature importance are irrelevant to the model. Following are examples of model interpretation methods.

Feature importance from the tree ensemble

Feature importances can be measured directly during model fitting thanks to the splitting error criterion. Accordingly, the Gini importance corresponds to the average decrease in Gini impurity, Equation (1.7), across all splits in the forest for a given variable [25, 71]. In the same line, the feature importance can be measured by counting the number of splits using each variable or the number of time a variable is used to initiate trees, i.e., is the first split [215, 17]. A drawback of feature importances derived from the trees is their reliance on tree structure. The selection of a variable to create a new split depends on the sample subset determined by previous splits. Hence, such measures are susceptible to perturbations in trees. Since different tree structures can result in the same model quality, feature importances relying on the tree ensemble structure are unstable, i.e., they vary across models fitted on different data sets from the same distribution [142, 107, 215].

The permutation importance measures the decrease in accuracy due to randomizing the variable vector [25, 107, 215]. For variables associated with the response, wrong decision paths would be followed for predictions and so permutations will lead to a drop in accuracy. For irrelevant variables, the measured decrease in accuracy would only be a consequence from noise and random associations [107].

Decision trees are known to be noisy due to their greedy building mechanisms [96, 71, 25]. Hence, distinguishing irrelevant from relevant variables based on the “raw” tree ensemble can be difficult. Certain implementations suggest to evaluate the FDR by adding random variables to the set of variables used for model fitting [123] or to estimate the variable importance distribution expected by random by repeatedly permuting the response variable [8]. Nonetheless, such methods do not actively prevent attributing high importances to irrelevant features.

Opening of tree ensembles with rule ensembles

The rule ensemble theory proposes to decompose decision trees into decision paths, named *rules*, to create a surrogate model [72]. Since tree ensemble make predictions based on partitions, limiting their ability to learn linear correlations, linear terms are additionally included to the surrogate. Hence, a regularized linear model is fitted such that the response variable is predicted using weighted rules and regularized variables. Here, model fitting consists of computing (i) rule weights that minimize a loss function reflective of the model predictive accuracy, and (ii) variable regularization terms that minimize the number of additional linear terms.

Friedman and Popescu [72] propose to compute the feature importance with two means. First, the importance of each rule $r \in R$ is measured using its weight a_r in the model and *sample support* s_r, i.e., the fraction of samples that follow the rule. It is then uniformely distributed across all m_r variables participating to the rule. For each variable, the importance due to rules is averaged across all rules to which it participates to. Secondly, the linear participation of a variable x_j to the model is measured via its regularized term b_j and standard deviation. Finally, the feature importance of a variable corresponds to the sum of its averaged contribution to rules weighted by the rule importance and to its linear importance [72].

This methods presents several advantages: (i) the noise from the original forest can be decreased via regularization and (ii) complex variable interactions captured by trees, and subsequently rules, are accessible. However, the need for a surrogate adds computational cost and, although being accessible, high-order interactions are not resumed in a comprehensive manner. For instance, in the inTrees R-package [48] implementation of the rule ensemble framework, the linear model surrogate is replaced by a RF trained on rules directly (the linear terms from the original surrogate are ignored), and the rules in the RF with the best Gini importance are returned. Therefore, data analysis has to be performed through two RFs and the interpretation of the final RF is not clear. Furthermore, the package does not provide any tool to understand variable interactions and feature importance.

Quite similar to rule ensembles, Random Intersection Trees (RIT) [205] suggest to evaluate all high order variable interactions via rules. In Basu et al. [14], authors proposed to extend the RIT framework by first training RFs, extracting rules from decision trees, and pruning them with the RIT algorithm. Bootstrapping is employed to obtain a stable ensemble of rules, and variable co-occurrences across rules are then calculated. A caveat in this method is the non-utilization of the informative model accuracy for assessing variable importance and interaction importance.

Model interpretation via local explanations

Finally, tree ensembles can be interpreted by looking at the contribution of variables to the prediction of each sample [71, 140, 191]. The general idea is to measure the change in prediction due to a variable value. Hence, such methods do not rely on the model shape but on its predictions only.

The partial dependence plot (PDP) [71] and individual conditional expectation (ICE)plot [79] estimate changes due to a fixed variable by screening predictions made for samples with the same variable value, for each value taken by the variable in the data set. While the ICE plots for each observation the curve estimating the conditional change of predictions given the values of all other predictors, the PDP outputs their average.

The LIME framework estimates variable contributions by fitting simple models around samples [190]. For a fixed observation x, new observations are generated by perturbing predictor values of x, model predictions are obtained, and a new local model predicting the original predictions is fitted on the set of perturbed samples. The local model can be of any sort, e.g., a linear regression or decision tree. It is employed to assess the contributions of variables to the prediction of each observation.

Finally, Shapley values build on the game theory, such that features are considered to be players of a game resulting in the prediction values [207]. Accordingly, for each game played (prediction) Shapley values measure the contribution of each variable to the game. Similar to LIME, the contributions of all variables are evaluated at once per-sample, instead of individually across samples as PDP and ICE. Many methods have been proposed to estimate Shapley values [217], among which the SHapley Additive exPlanation (SHAP) [140] that has been optimized for tree ensembles [142]. SHAP measures the effect on predictions of adding each variable to the model given all combinations of other predictors, hence it evaluates the additive effects of variables for each observation. For this, changes in predictions are recorded for sequential addition of variables and the expected change in prediction due to each variable is then calculated. Since the variable introduction order influences the effect of variables due to variable interactions, SHAP are approximated in practice to reduce computational costs [143, 18]. SHAP can be generalized to pairs of variables to assess the SHAP interaction for each feature in a pair [143]. The total interaction effect for a pair $\{x_i, x_j\}$ is equal to the sum of the SHAP interaction value attributed to each variable, and the SHAP interaction value of x_i is equal to the difference in effect when x_j is considered, i.e., when it is already present in the model, versus when it is absent. A concern for applying SHAP to tree ensembles is the assumption that variables have additive effects although models are not additive, which may lead to biased estimations of variable effects and variable interactions [84].

Local explanations are not specifically designed for the overall understanding of models and consequently may not be possible to summarize across observations. Furthermore, the number of plots to inspect the association between single variables and the target is equal to p, and the number of variable interaction plots is greater than $(p^2 - p)/2$. For $p = 10$, the number of plots already reaches 40; hence, such methods are often inappropriate for microbiome studies where p is high.

1.4 Outline

In the present work, I aim to provide support to bioinformatic analyses of sequence data with tree ensemble ML models. To this end, I first introduce a tool I created for interpreting tree ensemble models and benchmark it against state-of-the-art methods for analyzing sequence data. Then, I illustrate how tree ensemble models, together with my new method, can be applied to microbiome data. For this, the presence of methanogens in human gut microbiota is taken as an example. Finally, I provide supporting evidence of a key finding of the bioinformatic analysis using *in vitro* experiments.

Chapters 2 and 3 are dedicated to detailing endoR, a method for interpreting tree ensemble models. The method builds upon existing theories to offer a user-friendly, accurate, and comprehensive tool to describe tree ensemble models. In particular, it produces reliable results, thanks to bootstrapping, and displays them as an interaction network that enables visualizing complex interactions captured by predictive models. My proposed workflow, i.e., fitting tree ensemble models and interpreting them with endoR, surpasses state-of-the-art methods for analyzing sequence data.

The predominant human gut methanogen *M. smithii*, predominant in terms of relative abundance and prevalence, has been correlated to various host phenotypes. However, no thorough investigation focusing on *M. smithii* across human populations and using an appropriate data analysis workflow has been conducted to date. Chapter 4 presents results from such analysis, with a meta-analysis of human gut metagenomes performed using tree ensemble models and endoR.

Finally, in Chapter 5, I explore a main finding of my meta-analysis using *in vitro* experiments. Results from co-cultures of *M. smithii* with gut bacteria confirm the association of the methanogen with specific bacteria identified in Chapter 4, here due to cross-feeding.

Chapter 2
The endoR framework to interpret tree ensemble machine learning models

I will start my results sections by presenting the theoretical framework of a method I developed to facilitate the interpretation of tree ensemble models. While this tool is not directly related to metagenomes nor methanogens, it will be used in Chapter 4 to process a model predicting the occurrence of *Methanobacteriaceae* in human guts using metagenomic data. It was designed especially for this purpose but is applicable to any tree ensemble model.

Parts of this chapter will be submitted in an article still in preparation and the following text is adapted from drafts of the article. I conceptualized and implemented the whole method in R. Dr Niklas Pfister (University of Copenhagen, Denmark) provided valuable feedback on the method and participated to the mathematical writing. Dr Niklas Pfister and Dr Nicholas Youngblut reviewed and edited my original manuscript. All author contributions are detailed in Appendix A, Table A.1.

2.1 Introduction

The gut microbiome consists of microbial sub-communities exchanging and competing for resources, with the environment being shaped by the host [38, 193]. Relationships between the host and its gut microbiome are notably investigated using fecal microbiome sequencing.

For this, generalized linear models and covariance matrices adapted to support microbiome data are commonly employed for their ease of interpretation (e.g., SPIEC-EASI [124], MIMIX [87], mLDM [243] and kLDM [244]). Nonetheless, (i) generalized linear models do not capture complex interactions in predictive variables, and (ii) current implementations of covariance matrices are designed for sequence count data, hence limiting the exploration of microbiome in the light of host or study characteristics, e.g., country of origin or dataset in a meta-analysis. In consequence, tree ensemble models, i.e., random forests (RF) [25] and gradient boosted models (GBM) [71], are preferred to explore microbiome data [116, 115, 225]. RF models can successfully predict certain host phenotypes with microbial taxonomic and metabolic profiles generated from fecal metagenomes [171, 36, 182, 220, 65, 245]. However, tree ensembles are challenging to interpret due to their complexity.

Tree ensemble models are made of decision trees, with each tree denoting how groups of features partition samples by the prediction target [71, 25]. For example, a decision tree may segregate diseased from healthy individuals based on their high relative abundances of microbes A and B, but low relative abundances of microbe C. RF models build decision trees on random bootstrap resamples of features and observations, thus resulting in often hundreds of differing decisions trees [25]. Generating large forests leads to high accuracies with less overfitting but greatly increases the complexity of model interpretation.

Many efforts have been made to improve the interpretability of tree ensemble models, either via procedures to select the most relevant features and consequently decrease the complexity of models due to data dimensionality [48, 172, 187, 123], or via the measure of feature importance [2, 205, 140, 14, 17]. Feature selection is part of the model fitting workflow and is not a model interpretation tool *per se*, while feature importance is the state-of-the-art method for model interpretation after model fitting. The feature importance describes the strength of variable contributions to model prediction accuracy, the most common measures being the Gini and permutation importances [25, 27]. Recently, new feature importances emerged in microbiome science. For instance, Ai et al. [2] utilized the tree structure of RF to identify microbes predictive of colorectal cancers. Alternatively, Gou et al. [86] used SHapley Additive exPlanations (SHAP) [140] to select microbiome features associated with type-2-diabetes. Since SHAP do not inform on variable interactions, authors then correlated important microbiome features to host genetics and risk factors using generalized models.

Shapley values build on game theory to measure the contribution of variables to the prediction of each observation [207] and can be estimated from any model with many methods [217], including the SHAP method [140]. As they generate local, per-observation interpretations, they generally do not address the problem of the global feature importance [217]. In addition, despite being applicable to variable interactions, interactions are often calculated only for pairs of variables due to computational cost and their interpretation can be challenging for high-dimensional data sets [143]. Furthermore, the SHAP framework assumes variables to have additive contributions, even though tree ensembles are not additive models, which can lead to biased estimations of feature interactions [84].

Since decision trees utilize several features for prediction, they can inform on variable interactions with regards to predictions. Compared to variables never used together for predictions, variables repeatedly found on a same tree and branch are more likely to be associated. However, as tree ensembles are generated with a greedy procedure, especially RF, unimportant variables may occur along decision paths. To remove noise and facilitate the interpretation of tree ensembles, Friedman and Popescu [72] propose to measure the importance of variables in decision paths via lasso regression and then use the linear model as a surrogate to the RF. The inTrees R-package [47] and random intersection tree algorithm [205] implement similar ideas of simplifying tree ensemble models to create a reduced set of decisions from a forest. However, they lack tools to further interpret decisions. For instance, variables often co-occurring in trees may be co-dependent with regards to predictions (e.g., high abundances of both microbes A and B are associated with diseased individuals). The randomForestExplainer R-package [109] measures variable interactions in this manner. However, noise is not removed from tree ensembles before measuring variables co-occurrences and the package does not offer tools to easily interpret results of models fitted on high-dimensional data.

Due to the limitations of existing tools for interpretation of tree ensembles, I developed endoR, a method for interpreting tree ensemble models. The framework relies on decisions extracted from a tree ensemble model to measure the association of features, and pairs of features, with the response variable. For this, endoR extracts all decisions from decision trees, simplifies them via pruning and assess their stability via bootstrapping. The contribution of variables and pairs of variables in terms of importance and influence on predictions are then calculated. The importance corresponds to the gain in predictive accuracy attributed to variables (or pairs of variables), while the influence measures the change in model prediction due to the inclusion of variables (or pair of variables). Multiple intelligible plots are finally created to enhance readability of feature and interaction importances and influences. In particular, an interaction network where nodes correspond to variables and edges to interactions between them is generated. The size of nodes and edges in the network is proportional to their importance and colors indicate of the influence.

2.2 Theoretical framework

2.2.1 Rules, decisions and decision ensembles

Let $\mathbf{x} = (x^1, \ldots, x^p) \in \mathbb{R}^p$ represent p real-valued predictor variables (numeric or factor variables), $y \in \mathbb{R}$ a response variable, and assume we have observed a sample of n observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^{p+1}$. Both regression, where y is assumed to be a continuous variable, and binary classification, where $y \in \{0, 1\}$, are considered. In the case of multi-class classification, the problem needs to be transformed into a binary problem, e.g., one class against all other classes, which is automatically performed by endoR. Each predictor variable x^j is assumed to be either a numerical variable or a $0-1$ encoded factor variable. In particular, multi-level factor variables are assumed to be encoded by individual dummy variables (my implementation automatically performs this reduction).

A *rule* is a function $r : \mathbb{R}^p \to \{0, 1\}$ of the form

$$r(\mathbf{x}) = \mathbb{1}_{\mathcal{X}_r}(\mathbf{x}) = \prod_{j=1}^{p} \mathbb{1}_{\mathcal{X}_r^j}(x^j),$$

where $\mathcal{X}_r = \mathcal{X}_r^1 \times \cdots \times \mathcal{X}_r^p \subseteq \mathbb{R}^p$. A *decision* is defined to be a tuple $D = \{r_D, \hat{y}_D\}$ consisting of a rule r_D and a constant *prediction* $\hat{y}_D$.

The prediction $\hat{y}_D$ should be thought of as a good approximation of y on the *sample support* $S_D := \{i \in \{1, \ldots, n\} \,|\, r_D(\mathbf{x}_i) = 1\}$, the subset of samples following the rule. For each decision, $\hat{y}_D$ is computed either as the mean of y on S_D, in the case of regression, or as the fraction of samples in the positive class (defined by the user) on S_D, in the case of classification.

Decisions are the building blocks of a large class of non-parametric machine learning models such as decision trees, random forests and boosted trees. These models combine many decisions to construct high-capacity prediction procedures. Any such model can therefore be seen as a collection of decisions $\mathcal{D} = \{D_1, \ldots, D_M\}$, which we call a *decision ensemble*, together with an appropriate method for aggregating the predictions [96].

For every subset of observations $S \subseteq \{1, \ldots, n\}$, the error function $\alpha(S, \cdot) : \mathbb{R} \to \mathbb{R}$ is defined either as the mean residual sum of squares in the case of regression, or by the mean misclassification error in the case of binary classification, formally,

$$\begin{aligned} \alpha(S, \hat{y}) &:= \frac{1}{|S|} \sum_{i \in S} (y_i - \hat{y})^2 \\ \text{or } \alpha(S, \hat{y}) &:= \frac{1}{|S|} \sum_{i \in S} \big(1 - (\hat{y})^{y_i} (1 - \hat{y})^{1 - y_i}\big), \end{aligned}$$

respectively.

For a fixed decision D and a variable x^j, or pair of variables $\{x^j, x^k\}$, the *complement decision* D^{c}_j, or $D^{\mathrm{c}}_{j,k}$, are defined to be the decisions resulting from modifying rule r_D to have the complement support for the variable x^j, or the pair of variables $\{x^j, x^k\}$ (Figure 2.1), i.e.,

$$\begin{aligned} r_{D^{\mathrm{c}}_j}(\mathbf{x}) &:= \mathbb{1}_{\mathbb{R} \setminus \mathcal{X}^j_{r_D}}(x^j) \prod_{k \neq j} \mathbb{1}_{\mathcal{X}^k_{r_D}}(x^k) \\ \text{or } r_{D^{\mathrm{c}}_{j,k}}(\mathbf{x}) &:= \mathbb{1}_{\mathbb{R} \setminus \mathcal{X}^j_{r_D}}(x^j) \mathbb{1}_{\mathbb{R} \setminus \mathcal{X}^k_{r_D}}(x^k) \prod_{l \notin \{j,k\}} \mathbb{1}_{\mathcal{X}^l_{r_D}}(x^l), \end{aligned}$$

respectively.

Additionally, decisions D^{rm}_j and $D^{\mathrm{rm}}_{j,k}$ are defined to be the decisions resulting from removing the variable x^j, or pair of variables $\{x^j, x^k\}$, from the rule r_D (Figure 2.1), i.e.,

$$r_{D_j^{\mathrm{rm}}}(\mathbf{x}) := \prod_{k \neq j} \mathbb{1}_{\mathcal{X}_{r_D}^k}(x^k) \qquad \text{and} \qquad r_{D_{j,k}^{\mathrm{rm}}}(\mathbf{x}) := \prod_{l \notin \{j,k\}} \mathbb{1}_{\mathcal{X}_{r_D}^l}(x^l),$$

respectively.

Finally, for a subset of variables $J \subset \{x^j, j \in \{1, \ldots, p\}\}$ the decision D_J^{pr} is defined to be the decision resulting from removing all variables not included in J from r_D, i.e.,

$$r_{D_J^{\mathrm{pr}}}(\mathbf{x}) := \prod_{k \in J} \mathbb{1}_{\mathcal{X}_{r_D}^k}(x^k).$$

The predictions $\hat{y}_{D_j^{\mathrm{c}}}$, $\hat{y}_{D_{j,k}^{\mathrm{c}}}$, $\hat{y}_{D_j^{\mathrm{rm}}}$, $\hat{y}_{D_{j,k}^{\mathrm{rm}}}$ and $\hat{y}_{D_J^{\mathrm{pr}}}$ are each updated based on the new rule.

For a variable x^j, we define the set of *active decisions* as $\mathcal{D}^j := \{D \in \mathcal{D} | \mathcal{X}_{r_D}^j \neq \mathbb{R}\}$, the subset of decisions which depend on x^j. Likewise, the set of active decisions of a pair of variables $\{x^j, x^k\}$ is defined as $\mathcal{D}^{j,k} := \mathcal{D}^j \cap \mathcal{D}^k$.

2.2.2 Decision importance

For a decision $D \in \mathcal{D}$, the *decision importance* is defined by

$$I_D := \Big(1 - \frac{\alpha(S_D, \hat{y}_D)}{\alpha(S_D, \bar{y})}\Big) \cdot |S_D|.$$

This quantifies the improvement of predicting y on the support S_D with $\hat{y}_D$ instead of with the full sample average $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$. It is weighted by the size of the decision's support.

For regression and binary classification, $\Big(1 - \frac{\alpha(S_D, \hat{y}_D)}{\alpha(S_D, \bar{y})}\Big)$ corresponds to the coefficient of determination (or R^2) [53] and Cohen's κ [37], respectively, computed on the subsample S_D. Hence, the decision importance is a quality measure that incorporates both the support size and predictive performance of the decision.

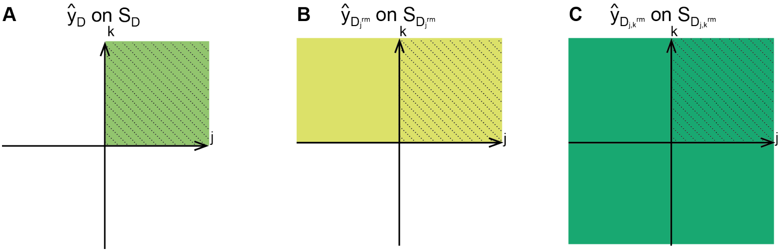

Figure 2.1: Prediction schemes used to calculate the participation of variables to decisions. Values that a decision D can take on variables j, k are heree represented. The support S_D is indicated by the stripped areas on A-C/, such as samples in the support of D all take positive values on j and k. B/ When variable j is removed from the rule r_D of D, the support $S_{D_j^{\mathrm{rm}}}$ is extended to samples taking negative values on j (colored area). C/ Similarly, a pair of variables $\{j, k\}$ is removed from r_D, samples in $S_{D_{j,k}^{\mathrm{rm}}}$ can take positive and negative values on j and k. For $S_{D_j^{\mathrm{rm}}}$ and $S_{D_{j,k}^{\mathrm{rm}}}$, $\hat{y}_{D_j^{\mathrm{rm}}}$ and $\hat{y}_{D_{j,k}^{\mathrm{rm}}}$ are calculated using all samples in $S_{D_j^{\mathrm{rm}}}$ and $S_{D_{j,k}^{\mathrm{rm}}}$, respectively. The decision-wise importance δ_D^j of j in D is calculated by comparing the error of $\hat{y}_{D_j^{\mathrm{rm}}}$ on S_D (B) versus the error of $\hat{y}_D$ on S_D (A). Similarly, to calculate the decision-wise importance of a pair of variables $\{j, k\}$ in a decision D, we compare the error from the decision not constraining values on j or k, with $\hat{y}_{D_{j,k}^{\mathrm{rm}}}$ on S_D (C) to the error of the decision with $\hat{y}_D$ on S_D (A).

2.2.3 Feature and interaction importance

To estimate the importance of a variable x^j within a decision D, predictive performance of two prediction schemes on the observations S_D are compared. The first prediction scheme uses the variable x^j to predict $\hat{y}_D$ on S_D, while the second prediction scheme does not use information about x^j and predicts $\hat{y}_{D_j^{\mathrm{rm}}}$ on S_D (Figure 2.1).

For a variable x^j, the *decision-wise feature importance* is defined as

$$\delta_D^j := \alpha(S_D, \hat{y}_{D_j^{\mathrm{rm}}}) - \alpha(S_D, \hat{y}_D),$$

the difference in predictive performance on S_D between $\hat{y}_D$ and $\hat{y}_{D_j^{\mathrm{rm}}}$.

For a pair of variables $\{x^j, x^k\}$, the *decision-wise interaction importance*

$$\delta_D^{j,k} := \sqrt{\delta_D^j \delta_D^k}$$

is the product of the decision-wise feature importances of x^j and x^k. I use the square root to ensure that the interaction importance remains on the same scale as the feature importance.

The *feature importance* and *interaction importance*,

$$F_j := \sum_{D \in \mathcal{D}} \delta_D^j I_D \qquad \text{and} \qquad F_{j,k} := \sum_{D \in \mathcal{D}} \delta_D^{j,k} I_D,$$

respectively, are then obtained by summing decision-wise feature and interaction importances over all decisions in $\mathcal{D}$ weighted by the decision importance. High values of the feature and interaction importances indicate that the variable, or pair of variables, participate a lot to important decisions.

2.2.4 Feature and interaction influence and direction

To understand how a single feature influences the prediction, one needs to understand whether a rule uses predominantly small or large values of that feature. For every decision D and variable x^j, the *direction indicator* $d_D^j \in \{-1, 1\}$

$$d_D^j := \begin{cases} 1 & \text{if } \frac{1}{|S_D|} \sum_{i \in S_D} x_i^j \geq \frac{1}{|S_{D_j^c}|} \sum_{i \in S_{D_j^c}} x_i^j \\ -1 & \text{if } \frac{1}{|S_D|} \sum_{i \in S_D} x_i^j < \frac{1}{|S_{D_j^c}|} \sum_{i \in S_{D_j^c}} x_i^j \end{cases}$$

expresses whether D predominantly uses small or large values of variable x^j.

The influence of a feature, or pair of features, on the prediction $\hat{y}_D$ of a decision is measured similarly to the feature importance, though actual predictions are now compared instead of errors of predictions on S_D.

For a variable x^j and pair of variables $\{x^j, x^k\}$, the *decision-wise feature influence* and *decision-wise interaction influence* are defined as

$$\gamma_D^j := d_D^j(\hat{y}_D - \hat{y}_{D_j^{\mathrm{rm}}}) \qquad \text{and} \qquad \gamma_D^{j,k} := \frac{d_D^j + d_D^k}{2}(\hat{y}_D - \hat{y}_{D_{j,k}^{\mathrm{rm}}}),$$

respectively. A large positive value of γ_D^j indicates that large values of x^j are positively associated with the response y on the support of the rule, while negative values of γ_D^j imply a negative association. Likewise, a large value of $\gamma_D^{j,k}$ indicates that large values of both $\{x^j, x^k\}$ are positively associated with y, and $\gamma_D^{j,k}$ is negative when small values of both $\{x^j, x^k\}$ are negatively associated with y. In addition, $\gamma_D^{j,k}$ is null when large values of x^j but small values of x^k are positively associated with y.

We assess the overall *feature influence* of a feature x^j, and *interaction influence* of pair of variables $\{x^j, x^k\}$, by averaging the decision-wise feature and interaction influences,

$$\Gamma_j := \frac{1}{\sum_{D\in\mathcal{D}^j} I_D} \sum_{D\in\mathcal{D}^j} \gamma_D^j I_D$$
$$\text{and } \Gamma_{j,k} := \frac{1}{\sum_{D\in\mathcal{D}^{j,k}} I_D} \sum_{D\in\mathcal{D}^{j,k}} \gamma_D^{j,k} I_D,$$

respectively.

For every pair of variables $\{x^j, x^k\}$,

$$\eta_{j,k} := \operatorname{sign}\Big(\sum_{D\in\mathcal{D}^{j,k}} (d_D^j \cdot d_D^k \cdot I_D)\Big)$$

records whether variables $\{x^j, x^k\}$ are each associated with y in the same direction across $D \in \mathcal{D}^{j,k}$. When both variables $\{x^j, x^k\}$ have large, or small, values associated with the response y, then $\eta_{j,k}$ is positive; and when large values of x^j are positively associated with y but small values are positively associated with y, then $\eta_{j,k}$ is negative. The later occurs when $\gamma_D^{j,k} = 0$.

2.2.5 Regularization of the decision ensemble

I propose several procedures to regularize the decision ensemble and so reduce the noise by including a simplicity bias. Procedures are recommended but optional.

Decision discretization Numerical predictors can be discretized based on their quantiles (e.g., into levels “Low”, “Medium” and “High”). All decisions containing discretized variables are then modified by replacing any numeric rule (e.g., “$x^j \leq t$”) by the best approximating rule which only uses the discretized variables (e.g., “x^j = ‘Low’”). Decisions consisting of the same rules are grouped, the multiplicity is recorded, i.e., how many decisions have been collapsed into the simplified decision) and the prediction, error, support and importances are re-computed based on the updated rule, and the decision importance is weighted by the decision multiplicity. Finally, the feature influence is computed for each level of discretized variables and the feature importance is calculated across all levels.

Decision pruning Pruning consists of removing variables from decisions that do not participate much to a decision, i.e., for which the difference in errors of the decision with and without the variable is low [47]. Comparison of errors can be performed using the absolute or relative difference in errors (absolute difference by default) [47]. Accordingly, the procedure looks for the smallest subset of variables J

with the lowest error, such as,

$$\alpha(S_{D_J^{\mathrm{pr}}}, \hat{y}_{D_J^{\mathrm{pr}}}) - \alpha(S_D, \hat{y}_D) \leq \theta \quad \text{or} \quad \frac{\alpha(S_{D_J^{\mathrm{pr}}}, \hat{y}_{D_J^{\mathrm{pr}}}) - \alpha(S_D, \hat{y}_D)}{\max\left(\alpha(S_D, \hat{y}_D), 10^{-6}\right)} \leq \theta, \tag{2.1}$$

with θ a user-specified threshold ($\theta = 0.05$ by default). If Equation (2.1) is not satisfied by any J, i.e., for all simplified decisions the differences in error are above the threshold, the original decision is returned. The prediction, error, support, importance and multiplicity are updated as described above.

Decision ensemble stability The decision ensemble will often be large and still include poorly predictive decisions. In addition, our method depends on the input data, which implies that for a new dataset drawn from the same distribution, the feature and interaction importance/influence might be slightly different. In consequence, decision filtering via bootstrapping and stability selection can be performed [153]. More explicitly, the decision ensemble $\mathcal{D}$ is first extracted from the predictive model and decisions are discretized if wanted. The entire method is then ran on B bootstrap resamples of the data, with pruning optionally performed and decisions' predictions, errors, supports and importances being calculated from the bootstrap data. A predictive model is not refitted, as this would be computationally too demanding. By default, bootstrapping is performed on $B = 10$ resamples of size $n/2$. The stable reduced final network is then obtained by adapting the stability selection procedure due to Meinshausen and Bühlmann [153]: all q most important decisions of each bootstrap are first aggregated, and those appearing in at least $\pi_{\mathrm{thr}} \cdot B$ of the resampled decision ensembles are then selected. For user-selected parameters $\alpha \in \mathbb{R}_{>0}$ and $\pi_{\mathrm{thr}} \in (0.5, 1]$ ($\alpha = 1$ and $\pi_{\mathrm{thr}} = 0.7$ by default), q is determined by

$$q = \left\lfloor \max\left\{1, \sqrt{(2\pi_{\mathrm{thr}} - 1) \cdot \alpha \cdot d}\right\} \right\rfloor,$$

where d is the average number of decisions across all resamples. For each decision in the stable decision ensemble, the decision-wise influence and importance are averaged across the resampled decision ensembles, and the influence and importance are re-computed as described above.

When bootstrapping is not conceivable, I propose to instead filter out all decisions with an importance below a given threshold λ_{imp}, selected by the user or using the following heuristic procedure

$$\lambda_{\text{imp}} := \underset{\lambda}{\operatorname{argmax}} \frac{|\mathcal{D}| - |\mathcal{D}(\lambda)|}{|\mathcal{D}(\lambda)|} \sum_{D \in \mathcal{D}(\lambda)} I_D,$$

where $\mathcal{D}(\lambda)$ is the set of decisions with $I_D \geq \lambda$.

Taxa aggregation I propose to take advantage of the feature importance to compare taxonomic ranks, aggregate them into the most relevant ones when possible, and thus facilitate interpretation. For this, taxonomic levels of a same branch are ranked according to their feature importance. First, if a taxonomic level has a lower rank than its coarser one, it is replaced in all decisions by the coarser level. In a second step, if a taxonomic level is represented by a unique finer level with a better rank, the coarser level is replaced by the finer one. Both steps are independent and can be performed separately of sequentially.

2.3 Implementation

I implemented the whole method described above, together with functions to visualize results, into an open source R-package available on GitHub (aruaud/endoR).

The main wrapper function of the endoR package takes as inputs (i) a predictive model fitted using the randomForest, ranger, gbm or XGBoost R-packages [129, 241, 88, 35], and (ii) data and a response variable on which to fit the decision ensemble, being the ones used to fit the model or not. Upon starting, all factor variables are transformed into dummy variables, and, in the case of multi-class classification, the problem is transformed into a binary problem according to the class defined by the user to focus on. All regularization steps, i.e., discretization, pruning, filtering and bootstrapping, are optional and parameters can be elected by the user. The current implementation was optimized using the data.table R-package, can be ran locally in parallel thanks to the parallel R-package, and bootstrapping can be further performed in parallel, locally or on a high-performance computing (HPC) environment, with the clusterMQ R-package.

I hereafter detail certain procedures that need additional clarifications.

2.3.1 Extraction of rules from predictive models

Decisions are extracted from tree-based models (randomForest, ranger, gbm and xgboost [129, 241, 88, 35]) using the inTrees R-package [47], with slight modifications. More specifically, given a tree-based model, rules are first extracted from all trees, or a subset of them, by following branches from the root down to the terminal node, e.g., for a tree composed of 4 terminal nodes, 4 decisions would be extracted.

From inTrees [47], I adapted the `treeVisit`, `extractRules`, and `ruleList2Exec` functions to be compatible with parallelization and return only full length rules. I also corrected the `Ranger2List` function that was deficient in inTrees [47].

2.3.2 Transformation of extracted rules

All multi-class factor predictive variables are then converted to $0 - 1$ encoded dummy variables. Extracted rules are then adjusted to be using only one class of each of the original multi-class factor variables and rules multiplicity is decreased accordingly. For instance, for a multiclass factor x^j taking values in $\{a, b, c\}$, three dummy variables would replace x^j and a rule such as "$x^j \in \{a, b\}$" would be transformed into two rules "$x^j_a = 1$" and "$x^j_b = 1$" with multiplicity equal to 0.5. In addition, the same procedure of rule splitting is applied for predictive factors that were already encoded as dummy variables for fitting the predicting model. Levels of multivariate variables are thus included only by their presence, later helping with the visualization and interpretation of networks.

2.3.3 Rule discretization

All, or a subset defined by the user, numeric variables are discretized based on their quantiles using the `discretizeVector` function from the inTrees R-package [47] that I adapted to accept missing values (NA). For each rule containing discretized variable, numeric thresholds are replaced by corresponding levels for which the majority of observations are included in the original sample support (Figure 2.2). Rules are then transformed as described in the above section to be based on only one level, and the multiplicity is updated.

2.3.4 Constructing the network

After regularization and computing all metrics, I propose to visualize the feature and interaction importance and influence in a network. In particular, nodes in the network correspond to single variables and edges to interactions between variables. More specifically, for every node $j \in \{1, \ldots, p\}$, we choose the node size and color in the following way:

- *node size:* feature importance F^j. Larger nodes correspond to more important variables;

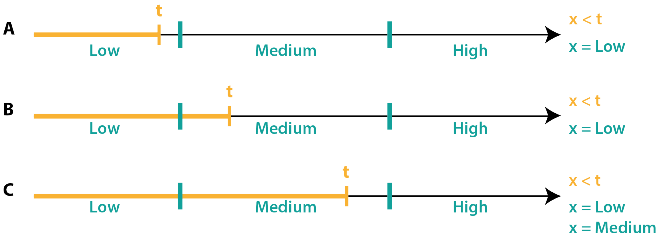

Figure 2.2: Discretization of variables and modification of rules. Simple example of the discretization of a uniformly distributed variable x into three levels. An original rule "$x < t$" (orange) is modified according to the number of observations in each level included in the sample support of the rule (new rule-s in green). B/ A minority of samples in the "Medium" level were included in the original sample support defined by "$x < t$", therefore the "Medium" level is not selected to make a new rule as in C/.

- *node color:* feature influence Γ_j, where the color interpolates from blue to orange (via white), with blue corresponding to small prediction values, white to prediction values close to the mean response variable across all samples, and orange to large prediction values.

Similarly, for every pair of nodes $\{j, k\} \in \{1, \ldots, p\}^2$, the edge between the two nodes is chosen as follows:

- *edge width:* interaction importance $F_{j,k}$. Thicker edges correspond to more important interactions;
- *edge color:* interaction influence $\Gamma_{j,k}$, with the same color scale than for nodes;
- *edge type:* interaction direction $\eta_{j,k}$. It is either a solid line if the pair of variables is on average used in the same direction in decisions, i.e., they are positively associated, and it is a dashed line otherwise.

The network object is created using the igraph and ggraph R-packages [39, 181], hence being compatible with the broadly employed ggplot2 R-package [237].

2.4 Discussion

The method I developed builds on (i) the tree ensemble theory [72] to open and simplify tree ensemble predictive models into decision ensembles, and (ii) Shapley values [72] to describe the role of features for predictions by decisions. Furthermore, it integrates regularization steps to prevent overfitting of the decision ensemble and facilitate the interpretation of results. Regularization is performed at two scales: decision-wise by directly simplifying decisions via pruning, and ensemble-wise by selecting the most important decisions via bootstrapping and stability selection. The method can be applied to any tree ensemble model and is therefore compatible with both regression and classification tasks, and with any type of data that can be predicted by the fitted predictive model. Thus, my method is appropriate for all sorts of problems investigated through tree ensemble model fitting.

I have implemented the method in an R-package, named endoR, available on GitHub. The current implementation offers means to visualize variable descriptors, i.e., the feature and interaction importances and influences, via intelligible plots to enhance the understanding of the final decision ensemble. In addition, the code is open source and outputs are readable by the base and ggplot2 R-packages [219, 237], hence enabling tuning by researchers for their own analyses. Furthermore, I optimized scripts and computation time with the data.table R-package [52] that utilizes multiple-threading for computations, thus decreasing computation wall-time. Within each bootstrap, local parallelization of tasks is possible using the parallel R-package, and bootstraps can be processed in parallel on HPC environments thanks to the clustermq R-package [200]. Additionally, certain tasks could be re-written in C++, notably using the Rcpp R-package [58], to accelerate computations. Accordingly, endoR is a user-friendly tool readily usable in R, a language routinely employed in biology, that can be effectively incorporated to data analysis workflows. Although not available in Python, a language also commonly used in biology, the similarities between R and Python would streamline the conversion of the endoR R-package into a Python-package.

In the next chapter, endoR will be applied to models fitted on simulated data and real metagenomes with artificial phenotypes to evaluate how it captures true associations from predictive models. It will also be benchmarked by comparing the accuracy of its feature descriptors, i.e., the feature and interaction importances and influences, to the accuracy of results from methods commonly used for metagenome analysis.

Chapter 3
Benchmarking of endoR

In continuity to Chapter 2, I will here benchmark the endoR method for interpreting tree ensemble models.

Similarly to Chapter 2, the following text has been adapted from an article in preparation. I produced all results and analyses. Dr Niklas Pfister provided much feedback on the design of simulations, their analyses, and on the mathematical writing. Dr Niklas Pfister and Dr Nicholas Youngblut reviewed and edited my original manuscript. All author contributions are detailed in Appendix A, Table A.1.

3.1 Introduction

Sequence data from gut microbiome have enabled exploring the relationships between human traits, e.g., body mass or diet, and microorganisms colonizing the last compartment where digestion ends. The gut microbiota is key to extract energy from undigested food such as resistant starch, and is in constant interaction with our immune system [29]. Consequently, efforts towards comprehensive studies of associations between gut inhabitants and host phenotypes have been undertaken [29].

A challenge in analysing sequence read data is that they are sparse, high-dimensional, and compositional [236, 124, 116]; hence, the need for appropriate tools to obtain confident, reproducible results. Although the objective is to recover as many true associations as possible, much care should be put on false positives (FP), particularly in biology where bioinformatic analyses are often used to orientate subsequent research [76, 188]. Therefore, the expected number of FP should be kept as low as possible to avoid expensive, time-consuming, and irrelevant experiments or investments. Classical statistical tests, e.g., Spearman's coefficient of correlation, and sparse covariance matrices, such as sparCC [70] and graphical lasso [69], are commonly used in microbiome science to infer microbe-microbe co-abundance patterns. Already five years ago, Weiss et al. [236] demonstrated that these methods produced unacceptable FP rates and were not better than random guesses in microbiome data configurations [236]; even though p-value correction and regularization were applied to prevent high false discovery (FDR) rates. Furthermore, sparse covariance matrices can only handle continuous, compositional, and zero-inflated variables, consequently precluding correction for covariates during association inference. This comment is relevant to classical pairwise tests too, which by nature, compare only two variables. Sparse covariance matrices also do not directly describe how microbe-microbe interactions associate with host phenotypes.

While methods to apply generalized models to microbiome data have been proposed [87, 243], a caveat of statistical modeling is the assumption that data follow a parametric model designed by researchers [26]. This has the advantage of producing simple and intelligible results, but is *a priori* incompatible with nature. The high FDR from sparse covariance matrices that are based on generalized models [236] gives evidence for generalized models to be inappropriate to sequence read data.

Complex models such as random forests (RF) can accurately predict host traits from sequence reads [135, 171, 36, 222, 182, 220, 65, 245]. Hence, they are recommended for investigating microbiome data although their interpretation is challenging [116, 115, 225]. Due to their complex structure, a measure of variable importances for making predictions is often the only information reported from these models [2, 85, 171, 220, 65, 182, 245, 36, 248], when other classical statistical analyses are not carried instead to characterize host-microbe and microbe-microbe associations [91, 59, 15, 110, 222]. Since ecological systems rely on complex relationships between their components [254, 29, 138], bioinformatic workflows should capture high-order interactions and outputs them in a clear manner. The SHapley Additive exPlanation (SHAP) method, implemented in Python and R, computes feature and interaction importances from models, including tree ensembles [140], and proved to be efficient on biomedical problems [141, 143]. However, calculations are made for each samples which complicates their understanding: SHAP values are to be examined for each pair of variables across the data set, i.e., one plot per pair, consequently rising the number of outputs to $(p^2 - p)/2$ (for $p = 10$ variables, 40 plots should be screened to detect interactions, and this rises to 190 when $p = 20$). Consequently, they may not be adapted for microbiome applications where data are highly dimensional and expected to have high-order interactions. Furthermore, in R, SHAP interaction values are implemented solely for gradient boosted models (GBM) [35] but not for RF.

In the previous chapter, I described endoR, a method I developed to enhance the interpretation of tree ensemble models (Figure 3.1). The method simplifies these predictive models into stable decision ensembles that preserve the predictive accuracy of the original models. A new measure of the feature importance is implemented and an additional measure describing the influence of features on predictions of the response variable is proposed. More specifically, the importance measures the gain in predictive accuracy attributed to a variable (or a pair of variables), while the influence measures how the inclusion of a variable (or a pair of variables) changes the models prediction. With the complexity of tree ensembles being reduced to a set of decisions, measures of importance and influence are easily extended to pairs of variables to seize variable interactions captured by the original model. Results are displayed as multiple intelligible plots to enhance readability of feature and interaction importances and influences (Figure 3.1 B and see Figure 3.3 for an application to metagenomes). Notably, endoR generates an interaction network where nodes correspond to variables and edges to interactions between them. Moreover, the sizes of nodes and edges are proportional to their importances, while their colors are indicative of their influence. Bootstrapping is readily integrated with stability selection to prevent overfitting and false discoveries. I implemented the full method into a user-friendly R-package that is computationally optimized, compatible with routinely used packages such as ggplot2 [237], and open source on GitHub.

I benchmarked my method on both fully simulated data sets and real metagenomes [176] with artificially generated phenotypes. In particular, I compared the endoR workflow with state-of-the-art procedures commonly used for analysing microbiome data. Results showed that endoR successfully extracts complex interactions from random forest (RF) models and performs better or comparable to existing methods. I then employed endoR on a metagenome dataset published by Qin et al. [185], in which the original study identified certain gut microbiome features to be associated with cirrhosis. From a single application of endoR, I was able to recover all major results of the original study and expand upon them by identifying additional oral-bacteria colonizing the gut of patients with cirrhosis and the depletion of bacteria associated with healthy microbiome [11].

My simulations demonstrate that tree ensemble models coupled with endoR are appropriate to analyze metagenome data. The proposed workflow is made of a few steps only, is user-friendly, produces accurate results with low FDR, and generates understandable outputs summarizing complex interactions from metagenome data.

3.2 Methods

3.2.1 Simulated data

Data I generated n independent observations of a random vector $(Y, K, V^1, \dots, V^{12})$ as follows. Let $V^1, \dots V^{12}$ be independent $\mathcal{N}(0.5, 1)$ distributed random predictive variables, let K, a multiclass predictor, be uniformly distributed over the categories $\{\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}\}$ and assume that the binary response variable Y is given by

$$\begin{aligned} Y = \big[\ & \mathrm{sign}(V^1V^2)\mathbb{1}_{K=\mathrm{a}} \ + \ \mathrm{sign}(V^3)\mathbb{1}_{K=\mathrm{b}} \ + \ \mathrm{sign}(V^4 + V^5)\mathbb{1}_{K=\mathrm{c}} \ + \\ & \mathrm{sign}(V^6 - V^7)\mathbb{1}_{K=\mathrm{d}}\ \big](2 \cdot \varepsilon - 1), \end{aligned} \tag{3.1}$$

where $\varepsilon \sim \mathrm{Bernouilli}(1 - r)$ adds noise by flipping the sign of Y with a probability of r. A set of correct decisions for this setting are given in Table 3.1.

I use this data generating mechanism as a toy model to evaluate endoR as the underlying mechanism is fully understood here.

Sets of simulations were performed with the following data parameters: n = 200, 1000 or 5000 samples and r = 0.05, 0.1 or 0.2 (with $n = 1000$ and $r = 0.05$ unless mentioned). Each set was replicated in 100 independent simulations (Figures 3.6 A-B and G-H), and a single replicate of the data with parameters $n = 1000$ and $r = 0.05$ is given in Figure 3.2 A-D.

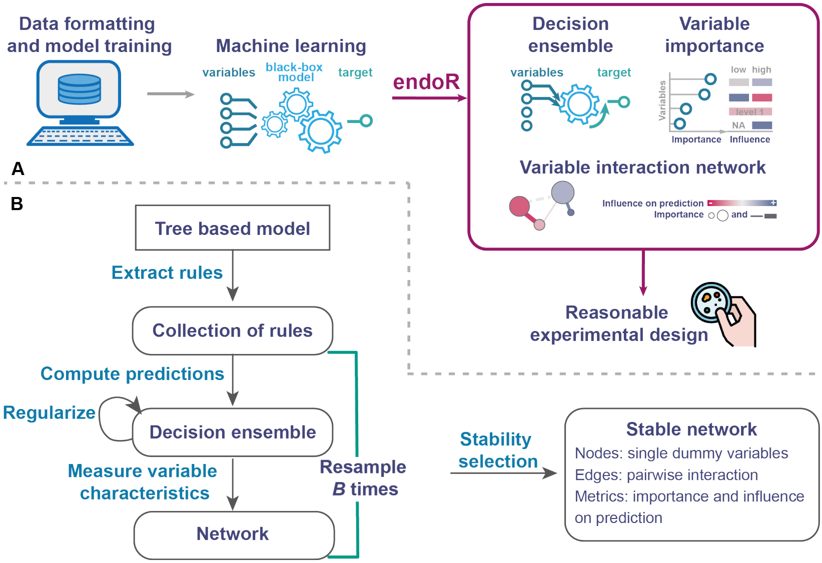

Figure 3.1: Description of the endoR workflow. A/ Overview of the workflow from data acquisition to the visualization of a network. endoR is applied to a fitted classification or regression tree ensemble model. The model is first simplified into a decision ensemble used to calculate feature importances and the influence on predictions. These metrics are displayed in a summary plot of individual variables and via a network for individual and pairs of variables. The network clearly illustrates the association between the response (target) and single or pairs of variables, in regards to feature importance and influence. Thus, if the influence of a variable depends on other variables, this will be visible on the network via edges between nodes. B/ Steps taken by endoR to generate a stable network. Regularization is optional and consists of simplifying decisions and the decision ensemble to reduce noise. The procedure can be repeated on B bootstraps to select stable decisions prior to constructing the final network.

Evaluation of endoR An RF model was fitted on each data set using the randomForest R-package with default parameters [129], unless mentioned (Figures 3.6 E-F). Classifiers were then processed with

Decision rule	Target label
Group = ’a’ & V1>0 & V2>0	1
Group = ’a’ & V1≤0 & V2≤0	1
Group = ’a’ & V1>0 & V2≤0	-1
Group = ’a’ & V1≤0 & V2>0	-1
Group = ’b’ & V3>0	1
Group = ’b’ & V3≤0	-1
Group = ’c’ & V4>0 & V5>0	1
Group = ’c’ & V4≤0 & V5≤0	-1
Group = ’d’ & V6≤0 & V7>0	-1
Group = ’d’ & V6>0 & V7≤0	1

Table 3.1: Predetermined decision rules based on the target equation from the simulated datasets.

endoR using default parameters, unless mentioned, and on $B = 100$ bootstrap resamples with $\alpha = 20$ (Figure 3.2 G-H) or $B = 10$ with $\alpha = 5$ (Figure 3.4 A-B and E-G). The average accuracy of the RF fitted on data in Figure 3.2 A-D was estimated on 10 cross-validation (CV) $0.7 - 0.3$ train-test. The accuracy of models fitted on all data is reported otherwise (Figure 3.6 A-B and E-F).

3.2.2 Artificial phenotypes

To assess the performance of endoR under more realistic microbiome conditions, I additionally evaluated it on a real metagenomic dataset with simulated response variables. Therefore, while the predictors are real data derived from metagenomes, I artificially constructed phenotypic groups and response variable to known ground truth of the underlying model. These data sets will be referred to as *artificial phenotypes*.

Metagenomes Data consisted of a subset of the metagenomes used in Youngblut et al. [246], so that samples with the following reported information were removed: i) samples from rectal swabs; ii) samples from individuals suffering from mumps, coeliac disease, gestational diabetes, cholera or with high relative abundances of *Vibrio cholerae*, infected by shiga toxin-producing *Escherichia coli* or cytomegalovirus; iii) samples with less than a million of sequence reads; iv) samples with missing age information. In total, metagenomes from 2147 samples from 19 studies and 23 countries were gathered. Only families, genera and species with a prevalence above 25 % relative abundances were included ($p = 520$ taxa).

Artificial phenotypes A multiclass phenotypic variable K, uniformely distributed either over the categories $\{a, b, c, d\}$ (Figure 3.3 A-D) or $\{a, b, c\}$, was constructed. Within each group, combinations of randomly picked taxa with a prevalence higher than 50 % were used to determine the sign of the response variable Y (for the replicate in Figure 3.3 A-D, see Table 3.2 and Figure 3.3 A-E). Noise was added by changing the group label with a probability r, such as the new label was drawn uniformly from all other groups and an additional irrelevant one.

Evaluation of endoR Predictive tree ensemble models were fitted as described in the following section 3.2.4. Each model was processed with endoR using default parameters and $\alpha = 10$ (Figure 3.4 C-D, H-J, and 3.6 C-D). For the replicate in Figure 3.3, numeric variables were discretized into 3 categories and $B = 100$ bootstrap resamples were performed with $\alpha = 5$ for stability selection.

3.2.3 Cirrhosis dataset

Finally, I evaluated endoR on previously published gut microbiome data used to predict cirrhosis.

Decision rule	Target label[a]
Group = ’a’ & $t_1 > 0$ & $t_2 > 0$	1
Group = ’a’ & $t_1 \leq 0$	-1
Group = ’a’ & $t_2 \leq 0$	-1
Group = ’b’ & $t_3 > 0.1$ & $t_4 > 0.003$	1
Group = ’b’ & $t_3 \leq 0.1$	-1
Group = ’b’ & $t_4 \leq 0.003$	-1
Group = ’c’ & $t_5 > 0$ & $t_6 > 0$ & $t_7 > 0.01$	1
Group = ’c’ & $t_7 > 0.01$	1
Group = ’c’ & $t_5 > 0$ & $t_6 > 0$	1
Group = ’c’ & $t_5 \leq 0$ & $t_6 \leq 0$ & $t_7 \leq 0.01$	-1
Group = ’c’ & $t_5 \leq 0$ & $t_7 \leq 0.01$	-1
Group = ’c’ & $t_6 \leq 0$ & $t_7 \leq 0.01$	-1
Group = ’d’ & $t_8 \leq 3.98 \cdot 10^{-4}$ & $t_9 > 0$	1
Group = ’d’ & $t_8 > 3.98 \cdot 10^{-4}$	-1
Group = ’d’ & $t_9 \leq 0$	-1

Table 3.2: Predetermined decision rules based on the making of the artificial phenotypes.

Data The dataset was composed of 130 samples for which age, BMI and sex information were available (68 cirrhotic and 62 healthy individuals). Metadata and gut microbial taxonomic profiles generated from metagenomes by Qin et al. [185] were downloaded from the MLRepo (https://github.com/knights-lab/MLRepo, accessed on 27/01/2021). The dataset consisted of stools samples from which gDNA had been extracted and sequenced via an Illumina HiSeq sequencer, and taxonomically profiles had been obtained using BURST [4] and Prokaryotic RefSeq Genomes. The downloaded taxonomic profiles consisted of read counts for taxonomic levels not collapsed at coarser level (ie, if a read count had been assigned to the species level, the number of count of the genus was not indicated). Consequently, I calculated the true read counts for each taxonomic level by summing read counts of all finer levels. Relative abundances were then normalized per the total number of reads to obtain relative abundances. Taxa were filtered as described in Chaper 4, section 4.2.1.

Evaluation of endoR I compared feature selection (FS) procedures to determine which full model to interpret with endoR (section 3.2.4 and Table 3.3). The final model was processed with endoR using default parameters, discretization into 3 categories, $B = 100$ bootstrap resamples of size $3n/4$. The number of samples that could be predicted from decision ensembles generated with various α values was used to select $\alpha = 30$ (Figure 3.9 B).

3.2.4 Fitting of models on metagenome data

Due to the high-dimensionality of metagenomes, feature selection (FS) was performed before fitting an RF model with default parameter [129]. A boosted tree model was alternatively fitted instead of the RF using the XGBoost R-package (default parameters and nrounds = 10) [35]. The choice of the FS algorithm and parameters was determined using 10 CV with a $0.7 - 0.3$ train-test split of the data: the model that resulted in the highest average Cohen's κ was selected and a final full-model was then refitted to the entire data (Table 3.3).

The types of models considered for the metagenome experiments were the following:

- `randomForest` function from the randomForest R-package (no FS);
- subselect variables using the Boruta R-package (`Boruta` and `TentativeRoughFix` functions with default parameters) and then apply `randomForest` from the randomForest R-package;
- subselect variables using the gRRF algorithm from the gRRF R-package for values of γ between 0 and 1 and, for each set of

features selected with a different γ value, apply `randomForest` from the randomForest R-package;

- subselect variables using a modified version of the gRRF algorithm to take into account the taxonomy (see Chapter 4, section 4.2.2), for values of γ and of k between 0 and 1 and, for each set of features selected with a different (γ, k) couple, apply `randomForest` from the randomForest R-package.

The choice of the Boruta and gRRF algorithms was motivated by the ability of Boruta to select all relevant variables [123], hence most likely to include all correlated variables, and for the ability of gRRF to select only relevant and non-redundant variables [48]. I additionally modified the expression of the regularization term in the gRRF algorithm, to account for the hierarchical taxonomic structure in metagenomes (motivation and method are detailed in Chapter 4, section 4.2.2).

Data	Feature selection[a]	RF accuracy[b](%)	Cohen's κ[b]	N features[d]
Main replicate of simulated phenotype	None	68.69±0.83	0.36±0.02	525
	gRRF (γ=0.45)	**85.19±2.36**	**0.70±0.05**	**18**
	taxa-aware gRRF[c] (γ=0.25, $k = 0.25$)	73.06±3.19	0.44±0.07	75
	Boruta	77.62±1.44	0.54±0.03	91
Cirrhosis	None	83.42±4.48	0.67±0.09	926
	gRRF (γ=0.1)	86.58±4.88	0.73±0.10	46
	taxa-aware gRRF[c] (γ=0.9, $k = 1$)	**86.58±3.81**	**0.73±0.08**	**69**
	Boruta	85.53±4.68	0.71±0.09	37

Table 3.3: Cross-validation (CV) of feature selection and training of classifiers on metagenomic data.

[a] The best FS selection algorithm is indicated in bold.

[b] Average and standard deviations across CV repetitions.

[c] A range of γ and k were tested for parameter tuning, but for concision, only results for the γ and k resulting in the best model are reported.

[d] Number of selected features for the model fitted on all data.

3.2.5 Evaluation metrics

Simulated data The ground truth network was extrapolated from Equation (3.1) (Figure 3.2 E). The network constructed from the final decision ensemble by endoR (Figure 3.2 H) was compared to the truth to count true positive (TP), false positive (FP), and false negative (FN) nodes and edges.

Artificial phenotypes Ground truth networks were extrapolated from the procedures used to create the artificial phenotypes (for an example, see Table 3.2 for the artificial phenotype in Figure 3.3 A-F and ground truth network in Figure 3.3 G). Since the data set is made of real metagenomes, a deficit here was the lack of ground truth on associations among predictive variables, notably from the same taxonomic branch. Hence, to account for taxonomic relationships, I also extended the lists of true nodes and edges to include nodes and edges from related taxa. I counted as related taxa the direct coarser and finer ranks, and species from the same genus. Consequently, a node identified by endoR was counted as TP if it was in the ground truth network, or related to a node in the ground truth network. If both a true node and a related taxa were identified by endoR, the TP was counted only once to prevent inflating results. The same counting was performed for edges.

Metrics Classical metrics (accuracy, precision, recall) were calculated to evaluate networks generated by endoR. In addition, TP and FP were weighted by their feature or interaction importances (for nodes and edges, respectively) to calculate the weighted precision, and so estimate the magnitude of TP in the endoR results. Furthermore, TP/FP curves can be constructed by any procedure that can rank the compared objects. To do so with endoR, for a fixed α, I first ranked the top q decisions of each bootstrap according to their probability of being selected in the final stable decision ensemble, i.e., they were ranked by their number of occurrences across bootstraps. Networks were computed for each probability of decisions of being selected, and the probabilities of edges and nodes to be in networks were subsequently calculated. Edges and nodes were then ranked by these probabilities and TP/FP curves were constructed for endoR (Figures 3.4 A-D, 3.5 and 3.6 H). Curves were interpolated and averaged across repetitions.

3.2.6 Benchmarking

Comparison of endoR with other analysis methods

To evaluate endoR against state-of-the-art methods for microbiome analysis, the artificial phenotypes were processed with the following procedures and compared to random guessing using TP/FP curves. Except for Figure 3.7 A-B, curves were interpolated and averaged across artificial phenotypes.

Pairwise comparison I used a Wilcoxon-rank sum test to identify taxa ($p = 520$ taxa) enriched in samples labelled with one or the other target category, and a χ^2-test to assess whether the K group categories comprised more samples than expected from one or the other target category; p-values were adjusted using the Benjamini-Hochberg correction method. Variables were ranked by increasing adjusted $1-p$-values to build the TP/FP curves.

Covariance matrix Sparse covariance matrices are used in microbiome science to determine conditionally non-independent taxa and build correlation networks [69]. The comparison of networks computed for distinct sample groups allows to identify different associations in these groups. For instance, by comparing networks extrapolated for samples collected from environment A versus environment B one can to infer associations of variables specific to each habitat [70]. A drawback of covariance matrices is the exclusion of categorical variables from analysis. Consequently, the K group variable was excluded from analyses, samples were divided by artificial phenotype label, and relative abundances of the $p = 520$ taxa were used to build sub-networks for each label. Edges shared between the two sub-networks were filtered out. Methods implemented in the SpiecEasi R-package [124] were employed to estimate covariance matrices, i.e., the sparCC [70], Meinshausen and Bühlmann [152] (MB), and graphical lasso [69] algorithms. Edges were ranked by the square value of the matrices parameter for making the TP/FP curves. Note that due to the low accuracy of the MB method [152], its results are not shown in the Results section. Lower performance was expected as this method is a simpler approximation of the covariance matrix suggested by Friedman et al. [69].

SHAP and Gini importances from tree ensembles Finally, I compared endoR to methods for interpreting predictive models. The Gini importance [25, 27] and SHAP values [140] were extracted from the RF classifiers used in Figures 3.3 and 3.6, using the randomForest [129] and iBreakDown [18] R-packages, respectively. In particular, SHAP values were calculated on the default number of 25 random paths, estimations were averaged across random paths for each sample, and for each variable, the absolute SHAP values averaged across samples were finally used to rank variables for the TP/FP.

Implementations of SHAP for RF classifiers in R do not return interaction values (see the iBreakDown [18], iml [158] and the fastshap R-packages). Consequently, I additionally fitted XGBoost models [35] on the same sets of features selected during FS when fitting the RF models, with default parameters, nrounds = 10, and objective =

'binary:logistic'. SHAP values and SHAP interaction values were extracted using the xgboost [35] and SHAPforxgboost [133] R-packages. Models were processed with endoR with default parameters. Figures of SHAP values extracted from XGBoost models were created using the SHAPforxgboost R-package [133] (Appendix A, Figure A.3).

Note that nodes and edges were ranked by feature and interaction importances to build the TP/FP endoR curves for comparison with the state-of-the-art methods.

Random guessing Variables, or pairs of variables, were randomly drawn and sorted to build TP/FP curve. The process was repeated a 1000 times and averaged.

Computation time

I measured the computation time and memory needed by endoR and the `shap` function (iBreakDown R-package [18]) to process RF models fitted the artificial phenotype of Figure 3.3 (Figure 3.8 and Appendix A, Table A.2). Since SHAP values can be directly extracted from XGBoost models, hence not requiring any additional processing time, I focused on comparing endoR and SHAP for RF models. Runs were performed in triplicates for the measurement of the total CPU time and maximal virtual memory used at any time (Appendix A, Table A.2), and in 5 replicates for the wall-time.

The same RF model as in Figure 3.3 was processed with endoR and `shap` using different input sample sizes, $n = 500, 1000$ or 2000, and number of bootstraps, $B = 1, 10, 20, 40$ (Figure 3.8 C-F). Furthermore, I increased the number of variables used in the predictive model by including non-selected features and fitting a new RF model via the randomForest R-package with default parameters [129].

Finally, the original model with 18 features and $n = 2147$ samples was processed with endoR and `shap`, with parallelization of calculations across 4 or 10 workers (controlled by the parallel R-package). For endoR, bootstraps were also allowed to be run individually in parallel using the clustermq R-package [200] (option clustermq.scheduler = ‘multiprocessor’). Wall-times were measured from runs on a machine equipped with Intel(R) Xeon(R) E5-4620 v4 @ 2.10GHz CPUs (80 CPUs in total).

3.3 Results

3.3.1 endoR recovers meaningful networks and metrics

Two sets of data were generated to evaluate endoR results. The first data sets were fully simulated so that predictors were independent from each other, normally distributed, and all associations of variables that could predict the target were known (Figure 3.2). For a more realistic configuration, I additionally used published human gut metagenomes comprising 2147 samples [176] and constructed artificial phenotypes from relative abundances of taxa generated from these metagenomes (Figure 3.3). Hence in the artificial phenotypes data sets, predictive variables were non-independent and consequently, not all predictive associations were known.

For both sets, different types of variable interactions were simulated to create the response variable. Observations were separated into four groups using a multiclass variable, and a binary response variable was computed via combinations of a group level and 1-3 of the continuous features (Figures 3.2 A-D and 3.3A-D, and Methods). I introduced noise in the fully simulated data by randomizing the target ($r = 0.05$), and in the artificial phenotype data by randomizing the group levels ($r = 0.05$). Here, I present results for one replicate of each configuration for which RF classifiers were trained and processed with endoR on

$B = 100$ bootstraps with $\alpha = 20$ for the simulated data and $\alpha = 5$ for the artificial phenotype (Figures 3.2 F-H and 3.3 G-J).

The RF accuracy and Cohen's κ, averaged across 10 cross-validations, were of 80.47±1.39 and 0.56±0.03 for the simulated data replicate, and of 85.19±2.36 and 0.70±0.05 for the artificial phenotype (Table 3.3).

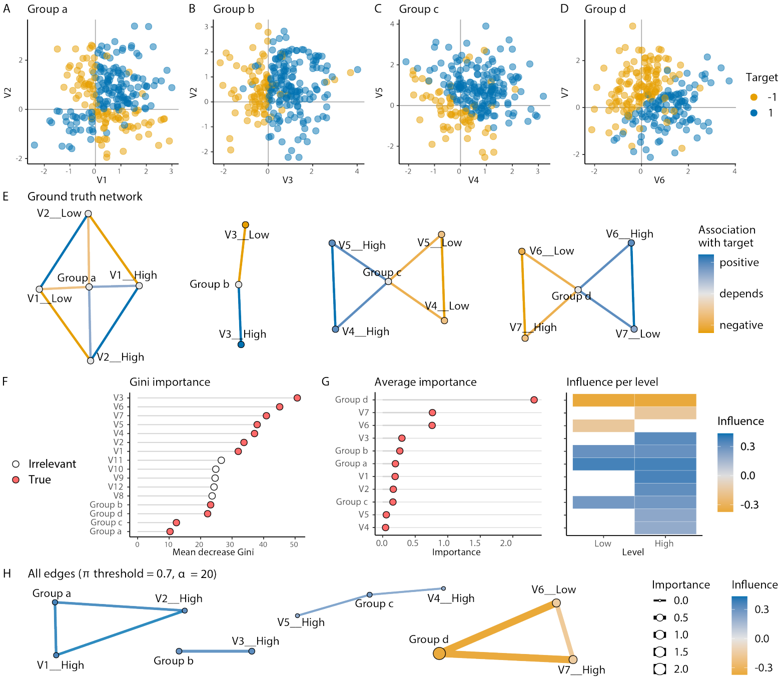

Figure 3.2: endoR captures interactions predictive of a response variable from a random forest fitted on simulated data. A-D/ Four groups of samples (labelled a-d) were generated so that for each group, the target takes the value '1' or '-1' according to a combination of variables (e.g., V1 and V2 for Group a). F/ Feature importance (mean decrease in Gini impurity) based on the fitted random forest. G/ Feature importance (summed across discretized levels) and feature influences for each discretized level computed by endoR. H/ endoR network produced from the RF model.

Consistent results were obtained across the two configurations. The feature importance measured by endoR better discriminated true predictive features from irrelevant ones compared to the commonly used Gini importance (Figures 3.2 F-G and 3.3 G-H). Furthermore, stable networks generated by endoR were coherent with ground truth networks, as they captured the interactions between numeric variables and groups (precision = 1 and recall = 0.90 for the simulated data set, and weighted precision = 0.57 and recall = 0.95 for the artificial phenotype; Figures 3.2 I,H and 3.3F,I-J). Finally, associations between the target and (i) each variable were conveniently displayed thanks to the feature importance and influence plot (Figures 3.2 G and 3.3 I-J), and (ii) variable interactions were easy to interpret thanks to the network (Figure 3.2 H and Figure 3.3 I-J).

3.3.2 endoR is robust to hyperparameters

I repeated the data and artificial phenotype simulations with different endoR hyperparameters to evaluate how the accuracy of endoR varies across differing conditions (Figure 3.4).

A sensitivity-precision trade-off is introduced by α

First, I explored the effect of α, which determines the expected number of wrong decisions selected by endoR after bootstrapping. As expected, the number of selected decisions increased with α (Figure 3.4 E,H). Accordingly, the number of TP and FP edges identified by endoR also increased (Figure 3.4 A,C,F and I). Nonetheless, the probability of true edges to be identified was still higher than the one of false edges, resulting in TP edges being first identified and having the highest importance in the stable networks (Figure 3.4 A,C,G and J). Hence, my results indicated a sensitivity-precision trade-off to consider when setting α and my general advice, especially for metagenomes, is to set α high enough

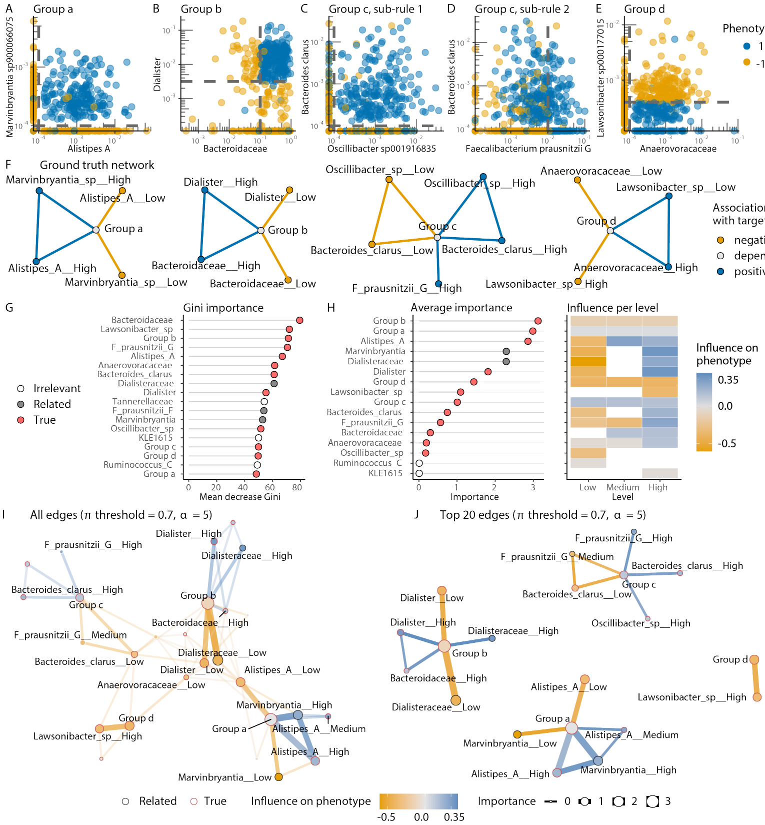

Figure 3.3: endoR captures interactions predictive of an artificial phenotype from an RF fitted on real metagenomes. A-E/ The artificial phenotype was computed from a group variable and randomly chosen microbial abundances. Dashed lines: thresholds used to make the phenotype. C-D/ Group c: the phenotype was built with an 'OR' rule. G-H/ Related: features related to the 'True' ones used to make the phenotype. I/ Full endoR network; only the 20 features with the highest feature importance labelled; the edge transparency is inversely proportional to the importance. J/ 20 edges with the highest interaction importance.

to recover a reasonable number of stable decisions and to particularly focus later interpretations on edges with the best importances.

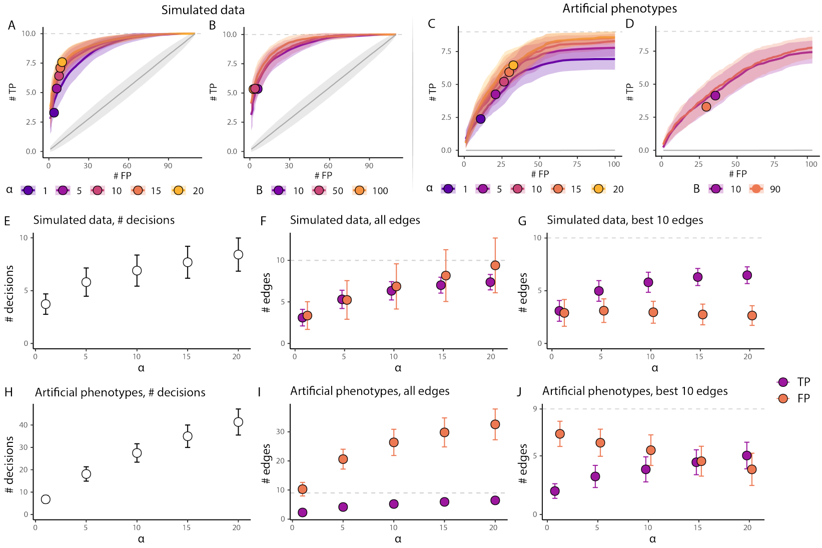

Figure 3.4: endoR is robust to hyperparameters. A-J/ Average number (#, lines or points) and standard deviations (shaded areas or bars) across simulations of identified true positive (TP) and false positive (FP) edges or of decisions in the stable decision ensembles; dashed lines: true number of edges. 100 simulated data sets ($n = 1000$) and 50 artificial phenotypes were generated per hyperparameter value ($r = 0.05$). A, C, E-J/ Effect of α, the expected number of false decisions in the stable decision ensemble ($B = 10$). B, D/ Effect of B, the number of bootstrap resamples ($\alpha = 5$ or 10 for the simulated data and artificial phenotypes, respectively). A-D/ Curves extrapolated for each simulation from the probabilities of decisions to be selected in the stable network; grey: results expected by random; traced points: TP and FP in the stable decision ensembles.

The stability selection procedure is robust to changes of B

Varying the number of bootstrap resamples between 10 and 90 did not affect the precision and sensitivity of endoR (Figure 3.4 B and D). This consistency in results suggests that on average, endoR results are similar for different number of bootstraps and that my stability selection procedure is efficient at discriminating relevant decisions. However, increasing the number of bootstraps helps obtaining steady decision ensembles for a same task processed with different random bootstrap resamples. This slight decrease in variance given higher number of bootstraps is exemplified in Figure 3.5 where I repeatedly processed replicates of the artificial phenotypes using distinct bootstrap resamples, for $B = 10$ or 100 bootstraps each time. Therefore, although endoR outputs similar results regardless of the number of bootstraps, those results are more likely to be closer to the expected average results with higher number of resamples. I thus recommend setting B as high as possible.

endoR gains in accuracy with better predictive models

Since endoR interprets tree-ensemble models, I then proceeded with evaluating the influence of input models on the accuracy of endoR. Assessment was performed using either or both simulated data sets and artificial phenotypes. The model accuracy was altered via (i) the model complexity through the number of trees in the RF (Figure 3.6 E-F), (ii) the noise in data by varying r, the probability of observations to be labelled with the wrong target, for the simulated data, or group category, for the artificial phenotypes (Figure 3.6 A-D), and (iii) the number of observations in data (Figure 3.6 G-H). The higher the number of trees in the forest, the lower the noise or the higher the number of samples, the higher were models' accuracy (Figure 3.6).

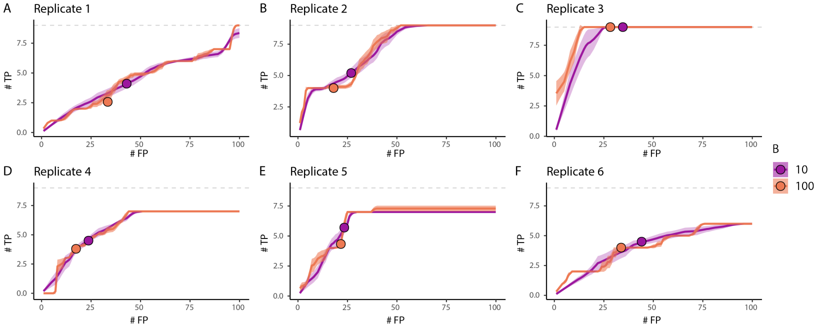

Figure 3.5: The variance of endoR results decreases with higher number of bootstraps. Replicates of artificial phenotypes were processed 10 times with $B = 10$ or 100 bootstraps resamples. Curves show the average number (#) of identified true positive (TP) and false positive (FP) edges according to edges probabilities of being selected in the stable decision ensemble. Curves were interpolated for each technical replicate, and the average (line) and standard deviation (shaded area) across number of bootstraps are displayed. Traced points: average number of TP and FP in the stable ensembles returned by endoR for $\pi = 0.7$ and $\alpha = 10$.

Taken together, results consistently showed that the performance of endoR depends on the quality of the input model (Figure 3.6). In particular, the weighted precision consistently benefited from improving predictive performances of forests (i.e., the accuracy or Cohen's κ), such that it increased with predictive performances of input models (Figure 3.6 A-F). Furthermore, the variance of the weighted precision across data sets decreased with increased predictive performances, meaning that although endoR produces rather precise networks on average, the probability of obtaining a precise network increases with the input model accuracy. The effect of model and data parameters on network recall was not as homogeneous as on the precision. For instance, although the recall greatly improved with lower noises for the simulated data (Figure 3.6 A), it barely increased for the artificial phenotypes (Figure 3.6 C).

Finally, while increasing the number observations for model training increased models' accuracies and so the weighted precision (Figure 3.6 G), it also allowed endoR to generate stable decision ensembles. Indeed, endoR produced a network via stability selection, with $B = 10$ and $\alpha = 5$, for 62 % of simulated data with $n = 200$ observations, whereas networks were produced for all simulations with $n \geq 800$. Nonetheless, for 45 % of all simulated data with $n = 200$, the precision of the network was equal to 1 (Figure 3.6 G-H). Therefore, for small sample sizes, endoR produces scarce but accurate networks, despite possibly not reaching a stable network. Furthermore, as the sample size increases, more complex and accurate networks can be generated by the method.

3.3.3 endoR surpasses the state-of-the-art for metagenome data analysis

I utilized the real metagenome data with artificial phenotypes to contrast the performance of endoR with the state-of-the-art (Figure 3.7). I processed each replicate of the artificial phenotype data sets with (i) a classical pairwise statistical analysis using the non-parametric statistical Wilcoxon rank-sum and χ^2 tests, (ii) sparse covariance matrices computed with the sparCC [70] and graphical lasso (gLASSO) [69] methods, (iii) RF classifiers fitted using the randomForest R-package [129], from which the Gini importance [25, 27] and SHAP values [140] were extracted, and that I further processed with endoR, and (iv) gradient boosted models fitted using the xgboost R-package [35] from which SHAP values were extracted and that I also further processed with endoR. For each replicate, single variables identified as associated with the artificial phenotype by SHAP (from RF and XGBoost models), Gini importance, and pairwise statistical tests

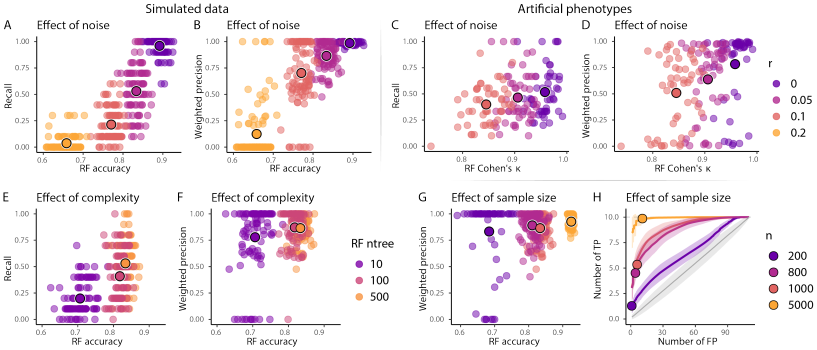

Figure 3.6: The accuracy of endoR increases with the accuracy of input models. A-B, E-H/ 100 simulated data sets were generated for each assessed parameter value, with $n = 1000$ and $r = 0.05$, as default parameters; models were fitted with ntrees = 500 as default, and were processed with endoR with $\alpha = 5$ and $B = 10$. C-D/ RF predictive of 50 artificial phenotypes were fitted with ntrees = 500, and were processed with endoR using $\alpha = 10$ and $B = 10$. A-H/ Large traced points: average across sets for a given parameter value. H/ Number of identified true positive (TP) and false positive (FP) edges according to edges probabilities of being selected in the stable decision ensemble. Curves were interpolated for each simulation, and the average (line) and standard deviation (shaded area) are displayed. The average number of TP and FP expected by random, and standard deviations, are shown in grey; dashed grey lines: actual number of TP.

were ranked by methods' output parameters and compared with the nodes identified by endoR from the RF and XGBoost models. Furthermore, the same procedure was carried for pairs of variables identified by SHAP (from the XGBoost model only as no method to compute SHAP interaction values from RF models is currently available in R), gLASSO, sparCC, and endoR.

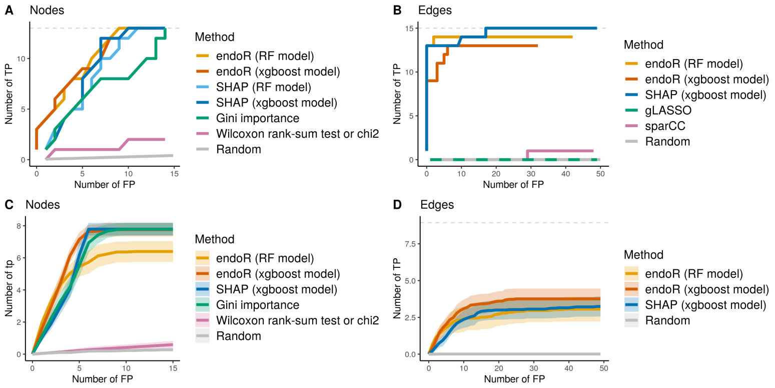

Figure 3.7: endoR is better or as good as state-of-the-art methods at identifying true variables and pairs of variables predictive of artificial phenotypes. TP: true positives; FP: false positives; dashed grey lines: true number of TP; Random: results expected by random guessing. A-B/ Results for the single replicate presented in Figure 3.3. C-D/ Average (line) and standard deviation (area) across repetitions of artificial phenotypes. B/ 'gLASSO' and 'Random' lines are dotted due to their overlap.

endoR is better or as good as state-of-the-art methods at identifying true variables and pairs of variables predictive of artificial phenotypes

All methods that did not use a predictive model (i.e., non-parametric statistical tests, sparCC and gLASSO) performed poorly, with accuracies nearly equivalent to random guessing (Figure 3.7 A-C). The generally good performance of methods based on classifiers was high thanks to the FS step performed with gRRF [48] before fitting models (Figure 3.7 A and C). As in my previous assessments, the Gini importance was not as good as endoR for discriminating true from irrelevant variables (Figure 3.7 A and C). SHAP values and endoR feature importances extracted from the RF and XGBoost models were both very precise in their ranking of features (Figure 3.7 A and C). SHAP derived from the XGBoost model was slightly better than endoR at identifying true interactions on the single artificial phenotype replicate (Figure 3.7 B, the replicate presented in Figure 3.3 had a more complex artificial phenotype creation mechanism than the repetitions). However, interaction importances extracted from XGBoost models for the repetitions of artificial phenotypes were better ranked by endoR than SHAP (Figure 3.7 D). Note that on average, XGBoost models were more accurate than RF models. The average Cohen's κ across repetitions was 0.97±0.00 and 0.91±0.03 for XGBoost and RF models respectively (mean across 10 CV sets for each replicate averaged here). And, for the single replicate, Cohen's κ was 0.95±0.01 for the XGBoost model and 0.70±0.05 for the RF model (averaged across 10 CV sets).

endoR results are easier to interpret than SHAP's

SHAP estimates the contribution of each variable, or pair of variables, to the prediction of each sample [140]. Thus, SHAP values are commonly visualized per sample, for each variable [141, 143, 133]; examples of SHAP visualization plots for the single artificial phenotype (Figure 3.3) are provided in Appendix A, Figures A.2 and A.3. Similar to endoR, summary plots of the feature importance can be produced by SHAP to provide a global overview of associations between the response and predictive variables. However, contrary to endoR, no summary visualization tool is available for SHAP to facilitate the interpretation of variable interactions and their associations with the response variable. Variables interactions are plotted for each pair of variables and each sample, such that as p increases, the number of plots to inspect exponentially rises. Furthermore, higher interactions cannot be assessed with SHAP since plots are created for pairs of variables, whereas endoR networks enable estimating such interactions.

endoR better scales with high-dimensional data and computing resources than SHAP for random forest models

Finally, I compared the computation requirements of endoR and SHAP on artificial phenotypes. Since SHAP values are computed by the xgboost R-package [35] while fitting the model, the two methods were compared on RF models only. SHAP values were generated from RF using the `shap` function from the iBreakDown R-package [18]. EndoR was much faster than `shap` and, in particular, was only linearly affected by the dimensionality and sample size while `shap` CPU time exponentially increased with dimensionality and sample size (Figure 3.8 A and C). The `shap` function used less maximal virtual memory used at any time than endoR. As expected the CPU running time of endoR linearly increased with the number of bootstraps (Figure 3.8 E). However, since endoR can be highly parallelized, shorter computation wall-times were measured for endoR with either $B = 10$ or 25, compared to `shap` (Figure 3.8 G).

My evaluations on all simulated data sets and phenotypes showed that endoR is more accurate that most state-of-the-art, and as accurate as the SHAP method. Furthermore it surpasses SHAP at facilitating model interpretation. Finally, compared to SHAP for RF, it better scales with highly-dimensional data and large sample sizes in terms of computation performance. Taken together, these results validated endoR as a powerful tool for investigating metagenomes with tree-ensemble models.

3.3.4 endoR recapitulates in one analysis previously reported results

To illustrate the utility of endoR for microbiome studies, I applied my tool to a previously published gut microbiome dataset comprising

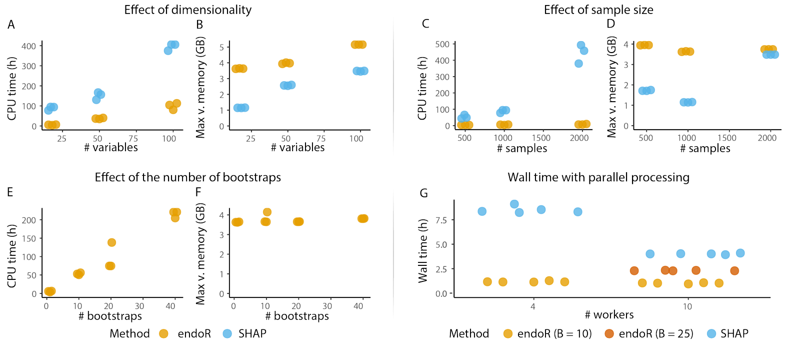

Figure 3.8: endoR is much faster to build an interaction network from a random forest classifier than the iBreakDown package calculating SHAP. A-F/ Total CPU time and maximal virtual memory used at any time (max v. memory) for three technical replicates of runs. By default, the artificial phenotype presented in Figure 3.3 was used with 18 variables and 1000 samples (see Methods), and endoR was ran on $B = 1$ bootstrap of size $n/2$. F/ Five technical replicates of runs on the main artificial phenotype (18 variables and 2147 observations), with parallelization of calculations across 4 or 10 workers and for endoR, bootstraps also allowed to be ran individually in parallel (see Methods).

patients diagnosed with cirrhosis versus healthy individuals [185]. The dataset included 130 Chinese subjects, among which 48 % were healthy, 35 % were women, with ages varying from 18 to 78 years old (45 years old on average), and BMI ranging from 16 to 29 kg.m^{-2} (22 kg.m^{-2} on average). In the original study, the authors tested for differences in microbial taxon relative abundance between cirrhosis and healthy patients. For this, they employed non-parametric statistical tests (i.e., Wilcoxon rank-sum test and Spearman's coefficient of correlation) with multiple testing corrections on both sequencing reads directly mapped to genome databases or grouped into metagenomic species before mapping. In addition, they constructed interaction networks by measuring correlations between taxa relative abundances with Spearman's coefficient of correlation.

I trained an RF classifier to predict the disease status (healthy or cirrhosis) of individuals based on their age, gender, BMI and relative abundances of gut microorganisms derived from sequencing reads directly mapped to a genome database (see Methods for details on model training and fitting). The model accuracy was on average across 10 CV train-test sets of 86.58±3.81 % and Cohen's κ was of 0.73±0.08 (Table 3.3). FS reduced the number of predictors from 926 to 85, to which I added back the age, BMI, gender and sequencing depth metadata to fit the final model on all samples. The model was then processed with endoR, on $B = 100$ bootstraps with default parameters and stability selection with $\alpha = 30$ (Figure 3.9 B). endoR identified 18 stable decisions that were using 23 predictors (Figure 3.9). Decisions, and feature and interaction importance and influence are given in Appendix A, Tables A.3 and A.4, respectively.

Many taxa used in the stable network were closely taxonomically related to taxa identified in the original study, with the same direction of association (Figure 3.9 A-B). I define 'closely related' as direct descendants or ancestors in the taxonomic hierarchy, as well as species from the same genus. As shown by Qin et al. [185], *Veillonella parvula*, *Megasphaera micronuciformis*, and members of the *Fusobacterium*, *Campylobacter*, *Lactobacillus*, *Streptococcus*, *Prevotella* genera, are enriched in individuals with cirrhosis; while members of *Eubacterium* genera, *Lachnospiraceae* and *Porphyromonadaceae* families are depleted (Figure 3.9 B). However, my results do not show a depletion in *Alistipes*, *Faecalibacterium praustnitzii*, *Coprococcus comes*, *Bacteroides* and *Ruminococcaceae* in individuals with cirrhosis.

Moreover, unlike Qin et al. [185], endoR did not identify any association between members of the *Veillonella* and *Campylobacter*, *Haemophilus* or *Fusobacterium* genera. These may have been spuri-

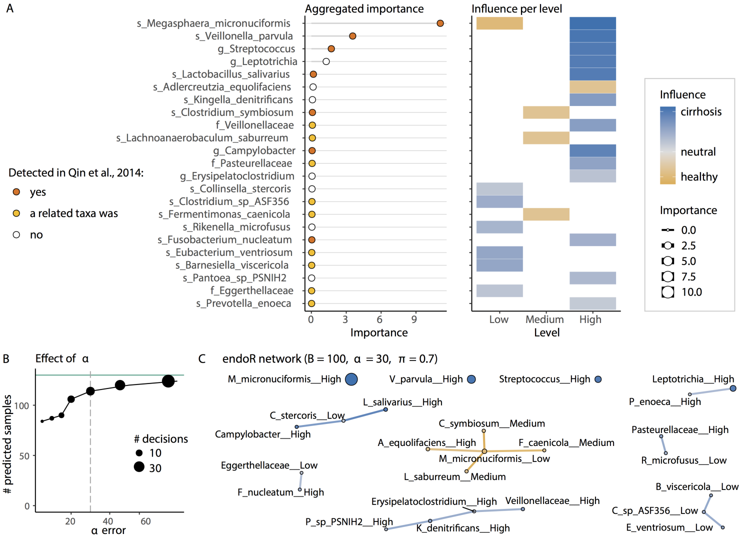

Figure 3.9: Exploration with endoR of gut microbiomes from healthy individuals versus patients diagnosed with cirrhosis. A/ Feature importance aggregated across each level of discretized variables and influence per-level as determined by endoR. Related: taxa directly coarser or finer, and species from the same genus, than taxa identified by Qin et al. [185]. White influence: the level was not used in any stable decision so that the influence could not be calculated. B/ Effect of α on the number (#) of decisions in the stable ensemble and samples that could be predicted by the ensemble. Green line: total number of samples in the dataset; grey line: chosen α to compute the stable decision ensemble (A,C). C/ Full network extracted from the stable decision ensembles. The boxed legend is shared for A and C.

ous associations detected in the original study due to a concomitant enrichment of these taxa in individuals with cirrhosis.

endoR identified two species to be the most discriminative between healthy and cirrhotic microbiomes: *Megasphaera micronuciformis* and *Veillonella parvula* (Figure 3.9 A). With the *Streptococcus* and *Lep-*

totrichia genera that had lower importances, they were the only ones to make decisions using a single variable, meaning that these taxa had a main effect (Appendix A, Table A.3). All other decisions in the stable ensemble predicted the health status of individuals using pairs of variables (Figure 3.9 C and Appendix A, Table A.3). Given the extremely low error on predictions of these decisions (on average, 0.02±0.01) and their support size (on average, 0.21±0.03), the interactions identified by endoR may also be relevant. Interestingly, endoR revealed an enrichment in the *Leptotrichia* genus in individuals with cirrhosis, notably associated with an enrichment in *Prevotella enoeca* (Figure 3.9 A and C). Those two taxa are part of the oral microbiome [49] and are enriched in patients with periodontal disease [131]. Periodontitis is more prevalent in individuals with alcohol-related cirrhosis, presumably due to a decrease in oral hygiene [89]. In addition, endoR identified in individuals with cirrhosis, an enrichment in members of the oral-taxon *Neisseriaceae* [49], notably of the *Kingella denitrificans* species. Altogether, these findings support the hypothesis of Qin et al. [185] of colonization of guts of patients with liver cirrhosis by oral commensals.

Furthermore, endoR distinguished an enrichment in *Adlercreutzia equolifaciens* in healthy individuals relative to individuals with cirrhosis (Figure 3.9 A). This finding is coherent with the previously observed depletion of *A. equolifaciens* in patients with primary sclerosing cholangitis, a condition that can lead to cirrhosis [11]. The species was shown to depleted while *M. micronucformis* was enriched, as reflected in the endoR network (Figure 3.9 C).

3.4 Discussion

I have shown that endoR, my tool for interpretation of tree-ensemble machine learning (ML) models, accurately captures the main variables and interactions of variables that predict a target of interest (e.g., disease status). To this end, endoR combines the conceptual backbone of rule ensemble theory [72] to simplify complex models, with game theory reasoning for the calculation of variable contributions [207]. It allows it to recover meaningful networks and feature importance. Furthermore, the method is proposed with regularization and stability selection to prevent overfittting of results [153]. My simulations showed that endoR is robust to hyperparameters, hence not demanding particular tuning.

EndoR processes tree ensemble models fitted on categorical and discrete features; hence, my method enables exploring not only microbial abundances but also metadata such as gender and age. Results from benchmarking showed that endoR was more accurate than state-of-the-art commonly used measures, i.e., Gini importance, non-parametric statistical tests, and sparse covariance matrices. It was as accurate as SHAP and, compared to SHAP, endoR had the benefit of providing intelligible outputs to facilitate interpretation, i.e., the feature importance and influence plot, and the interaction network. Notably, SHAP does not provide any summary output allowing to get an overall idea of variable, and interaction of variables, associations with the response. As microbiome studies include often more than hundreds of variables, an effective comprehension of models with SHAP is hindered by its lack of summary visualization tool. In particular, as the gut microbiome is a dynamic environment where microbes and host all interact with each other [254, 29, 166], the interaction network is critical for understanding the mechanisms by which microorganisms alter host phenotypes. Furthermore, in R and for RF models, endoR better scales to large datasets than SHAP in terms of computational time. Finally, endoR was the only method to readily integrate both regularization and bootstrapping, enabling to generate accurate results while preventing overfitting [153].

EndoR allowed me to easily explore differences in gut microbiota between healthy and cirrhotic individuals [185]. For this, I could use a unique analysis consisting of a FS step, the fitting of an accurate RF classifier and finally, the interpretation of the RF model with endoR. My findings confirmed results from the original study and reinforced them with new insights. Notably, I identified additional oral-bacteria enriched in guts of cirrhotic individuals, as well as a concomitant depletion in *A. equolifaciens*, a healthy-gut-associated microbe, and enrichment in *M. micronuciformis* in cirrhotic individuals compared to healthy ones. Nonetheless, the feature importance calculated by endoR highlighted the main effects in distinguishing cirrhotic from healthy individuals due to *M. micronuciformis* and *V. parvula.*

While I have shown that endoR is a powerful tool for ML model interpretation, the approach has limitations. Similar to all model interpretation methods, the accuracy of endoR is proportional to the input model accuracy. My extensive simulations provide a guideline on what accuracy to expect from endoR, depending on the accuracy of the ML model. Regardless, researchers should fit models of good quality prior to applying endoR or any interpretation method. Another general limitation of endoR is its specific design for tree ensemble models, making it incompatible with other algorithms. Nonetheless, as RF and GBM often outperform other algorithms when applied to microbiome data [225, 116], endoR is relevant to the field.

Due to filtering and bootstrapping, it may occur that the final decision ensemble cannot predict all observations, i.e., samples are not part of any decision support. In such cases, the α parameter can be increased until all samples belong to the support of at least one decision or bootstrapping can be performed on on larger data sets. Furthermore, non-predicted samples may be outliers requiring additional analyses or information to elucidate the mechanisms behind their response variable value.

Finally, although I primarily designed endoR to be compatible with omic data, my current implementation can be demanding in time and memory for complex models, e.g., large forests trained on data with $p > 100$ & $n > 1000$. Hence, I generally advise users to include FS in their model fitting protocol to decrease p, which will also most likely result in better predictive models. If memory issues arise, setting the number of categories for discretization to 2 (default) can solve the problem. Otherwise, the model complexity may need to be adjusted by performing a more constraining FS step, or fitting a smaller model, e.g., an RF with less trees.

Through extensive evaluations on simulated and real data, I have demonstrated that endoR, my method for interpreting tree-ensemble ML models, outperforms the state-of-the-art with regards to accuracy, robustness, and ease of interpretability. EndoR provides sharp insights into pairwise associations of features with the response variable, and produces a clear network to assess high-order interactions among variables. Tree-ensemble models are more frequently utilized to investigate relationships between microbiome sequence data and host phenotypes such as disease states. EndoR helps to unlock the mechanisms by which these black-box models make accurate predictions. Such insights are needed to move beyond predictive models and determine the dynamics underlying the modulation of host phenotypes by the gut microbiome and vice versa. Past the scope of this work, there is no restriction to applying endoR on ML models fitted on any problem, being biology-related or not.

Chapter 4

The occurrence of *Methanobacteriaceae* across a large population is predicted by relative abundances of a consortium of bacteria

This chapter will present results from analyses of human gut metagenomes to identify patterns of bacterial and metabolic pathway relative abundances associated with the presence of *Methanobacteriaceae*. Considerations related to data processing arising from previous chapters were applied here: (i) shotgun metagenomes from multiple human populations were used for analyses, (ii) as many metadata as possible were included to account for population and study biases, (iii) many taxonomic ranks were also included as limited prior knowledge of the important ones was available, (iv) tree-ensemble machine learning models were trained to predict the presence/absence of methanogens based on their compatibility with highly-dimensional and compositional data, and (v) taxa-aware feature selection (FS) was performed to decrease the number of predictive variables. Models were interpreted with endoR, the method I created and described in Chapter 2.

4.1 Introduction

The human gut methanogen *M. smithii* stands at the end of a trophic chain that starts with the host's food digestion along the gastrointestinal tract and continues in the intestine with the gut microbiome. Undigested carbohydrates and proteins are degraded and fermented by unique, or consortia of, bacteria [63, 64]. Notably, this process results in various products, including H_2, CO_2, and formate. *M. smithii* then transforms these substrates into CH_4, a gas excreted from the host [138, 196]. SRBs and acetogens can also use H_2, rendering them potential competitors of the methanogen. Fermentation substrates and pathways are specific to bacteria, such that the composition of the pool of undigested products will shape the microbiome composition [102, 193, 29]. Ultimately, this will influence the colonization ability of *M. smithii*.

In line with these interconnections, *M. smithii* is associated with carbohydrate-rich diets [31, 42, 147, 170] and H_2-producers fermenting distinct substrates, e.g., the cellulose-degrading *Ruminococcus* sp. [31] or members of the *Christensenellaceae* family [80, 94, 230, 114], which grow on simple sugars [159, 149, 253]. However, contrary to expectations based on competition for H_2, *M. smithii* and SRBs are positively correlated, probably owing to shared niche preferences [94]. By uptaking fermentation products, the methanogen is believed to promote fermentative pathways that produce methanogenesis substrates in the gut [196, 46, 120, 138]. By doing so, it would modify SCFA production, therefore influencing host metabolism, which would support its association with BMI [144, 9, 157, 80, 201, 106, 28, 150, 249, 228]. In addition, *M. smithii* has a slow generation time *in vitro*. Accordingly, its reported associations with slow transits and constipation have been suggested to result from a lower wash-out effect by digestive tracts [121, 238].

Although valuable, most studies reporting associations between human gut methanogens and their environment rely on statistical tests not designed for compositional data, as is the case for all studies cited in the previous paragraph. Colinearity between methanogens and bacterial relative abundances is commonly assessed using Spearman's or Pearson's correlation tests, and relative abundances in sample groups (e.g., healthy versus diseased) are frequently compared with a Kruskal-Wallis test or a Student t-test. Such methods only allow pairwise associations and may thus miss interactions between microorganisms. Furthermore, despite the growing use of shotgun metagenomes for microbiome investigations, some findings are supported by breath CH_4 measurement [31] or sequencing of the 16S rRNA gene [80, 94, 230, 114] to cite merely a few. Due to biases inherent in these detection methods, methanogens' occurrences and relative abundances may be underestimated in those datasets, potentially resulting in a lack of statistical power to infer associations.

As a whole, we have acquired disparate knowledge of *M. smithii*'s ecology but still lack global insights. To gain a broad picture of its interactions with the human gut microbiome, I performed a meta-analysis of gut metagenomes from 26 studies, representing individuals from 23 countries worldwide, using tree-ensemble models that can capture interactions between variables. Predictive models were interpreted using endoR and showed that *M. smithii*'s occurrence is highly associated with the presence of members of the CAG-138 family, order *Christensenellales*, specifically with the Phil-1 genus, as well as with members of the *Oscillispiraceae* family. In particular, high relative abundances of the glycolysis IV pathway coupled with low relative abundances of *Oscillispiraceae* spp. were indicative of an absence of methanogens in samples, and inversely, low relative abundances of the metabolic pathway coupled with high relative abundances of *Oscillispiraceae* spp. or *Christensenellales* spp. were indicative of methanogen presence. Moreover, host characteristics, such as body mass index (BMI) or enterotype, did not prove to be predictive of the presence of *Methanobacteriaceae*, suggesting that the microbiome composition primarily determines the ability of methanogens to colonize human guts. Taken together, my results provide new perspectives on the prevalence of methanogens in the human gut and their plausible interactions with members of the gut microbiome.

4.2 Methods

4.2.1 Data

curatedMetagenome database

Data used in this chapter were downloaded from the curatedMetagenomic database [176]. I included all samples from Youngblut et al. [246], except for samples meeting the following additional exclusion criteria: (i) samples from rectal swabs; (ii) from individuals older than 90 years old, with a BMI greater than 40 kg.m^{-2}, with any reported disease, or not part of control cohorts; (iii) samples from David et al., 2015 [40], due to the infection of all individuals with *Vibrio cholerae* or enterotoxigenic *Escherischia coli*; (iv) samples with less than a million sequence reads.

Information about sampled individuals comprised: country of origin, age, BMI, and whether the individual was from a westernized population. Here, westernization should be understood as a urban lifestyle with a diet composed of fewer carbohydrates and enriched in fat, sugar and animal products compared to rural populations [183, 42]. The dataset consisted of 2203 samples from 26 studies and 23 countries, among which 748 samples had complete gender, age, and BMI information (Appendix B, Tables B.2 and B.1).

Enterotype clustering

Enterotypes were determined as described in Arumugam et al. [10]: the Jensen–Shannon distance matrix was calculated from the relative abundances of genera using the ape [173] and phytools [189] R-packages, and partitioning around medoid was then performed with the cluster R-package [145].

Metabolic pathways formatting and filtering

Relative abundances of metabolic pathways were downloaded from the curatedMetagenomic database [176], where they had been obtained thanks to the HUMANn2 pipeline [67]. All engineered, unmapped and unintegrated pathways were removed. Furthermore, only relative abundances of pathways at the community level, i.e., calculated from all gene abundances in the sample, were considered for analysis. Accordingly, I removed all relative abundances calculated from species-level gene abundances, i.e., the abundances attributed to distinct species [67]. I additionally crossed the relative abundance and coverage of pathways. The HUMANn2 pipeline calculates a confidence score that indicates whether reactions of pathways with non-zero relative abundances are confidently detected. A pathway coverage of 0 means that although genes coding for proteins involved in this pathway were detected, not all reactions of the pathway were confidently mapped [67]. For this reason, for each sample and metabolic pathway, the relative abundance was replaced for 0 if the coverage was null. Finally, all pathways present in less than 25 % of samples were removed. A total of 117 metabolic pathways were included in analysis.

Taxa abundances filtering

Sequence reads were processed by Dr. Nicholas Youngblut for taxonomic profiling [246]. I performed multiple taxonomic filtering steps to reduce sparsity, taxonomic redundancy, and ultimately the number of variables.

Filtering of rare taxa I took a progressive approach to filter out rare taxa, with low average abundance or low prevalence. Hence, a taxa $t \in T$ of prevalence P_t and average abundance A_t was removed if:

$$P_t < A_t \cdot \beta_0 + \beta_1,$$

with

$$\begin{aligned}\beta_0 &\coloneqq \frac{\text{median}_{i \in \{1,\dots,T\}}(P_{t_i}) - P_{q1}}{A_{q1} - \text{median}_{i \in \{1,\dots,T\}}(A_{t_i})}\\ \beta_1 &\coloneqq P_{q1} - \beta_0 \cdot \text{median}_{i \in \{1,\dots,T\}}(A_{t_i}).\end{aligned}$$

P_{q1} and A_{q1} correspond to the prevalence and abundance quantile values of 25 % of all taxa. This continuous filtering allows me to keep taxa that are highly abundant in only a few samples, and conversely to keep taxa with high prevalence across samples but low abundances (Figure 4.1). This filtering was performed with pooled family, genus and species taxonomic ranks.

Filtering of correlated taxa To limit redundancy in relative abundances from taxonomic ranks of a same branch, I filtered out taxa that were significantly correlated to their direct coarser level [172]. A Spearman test was performed between the two taxa, and the finer one was removed if p-value < 0.05 and $\rho^2 \geq 0.95$. A total of 89 taxa were filtered out in this manner.

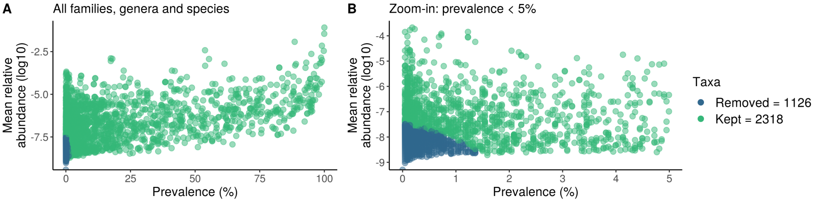

Figure 4.1: Mean relative abundances and prevalence of family, genus and species taxonomic levels in the metagenomic data.

4.2.2 Data analysis

General workflow to fit a model predicting the presence of *Methanobacteriaceae*

As DNA extraction protocol can alter archaeal DNA recovery, and so influence the relative abundance of methanogens in samples, I looked for associations between taxa and metabolic pathways relative abundances, and metadata, with methanogens' presence. Their occurrence was inferred from non-zero relative abundance of *Methanobacteriaceae*.

Random forests (RF) were employed for analyses as they are compatible with high-dimensional, compositional, sparse and correlated data [116, 115]. I used the ranger R-package [240] to fit RF models and account for data imbalance. To accomplish this, two strategies were tested: providing class weights to tune the learning process and obtain a cost-sensitive model (class.weight parameter), and providing sampling probabilities inversely proportional to classes' distribution (case.weights parameter). I also trained gradient boosted models using the XGBoost R-package [35], but these resulted in lower predictive performances (Appendix B, Table B.3). Thus, they were not further utilized.

Finally, models' performances were evaluated on 10 cross-validation (CV) 70-30 % of train-test sets. Model processes were fitted to training sets and predictive performance was measured using Cohen's κ on test sets.

For model selection, I restrained model complexity by taking into account the number of features used for fitting models and the number of trees in the forest. Let w_m be the scaled weight of model descriptors for each model $m \in M$, and M the set of all fitted models across model sequences and cross validation sets, then

$$w_m := 1 - \frac{\max\{n_T^i \cdot n_{FS}^i \,|\, i \in M\} - \left(n_T^m \cdot n_{FS}^m\right)}{\max\{n_T^i \cdot n_{FS}^i \,|\, i \in M\} - \min\{n_T^i \cdot n_{FS}^i \,|\, i \in M\}} \,,$$

with $w_m \in [0, 1]$, n_T the number of trees in the forest, and n_{FS} the number of selected features. The combination of FS parameters and number of trees that minimized Cohen's κ weighted by w_m was selected as being the most optimal with the best predictive performance but lowest complexity. This strategy was used to compare RF models trained on all observations, i.e., without gender, age, and BMI in the set of predictors, with taxonomic ranks ranging from the family to species.

Sets of predictors

I fitted models on different sets of predictors to reduce dimensionality. Since gender, age, and BMI were incomplete (Appendix B, Table B.1), I first assessed whether those variables were selected and used in models fitted on the 748 samples with complete information ($n_T = 500$ and cost-sensitive model). Otherwise, models were fitted on all samples without gender, age, and BMI. Included metadata were added in each model processing step, even if they were not selected during the FS. For taxonomic relative abundance, I first included taxonomic ranks from family to species, totaling 2206 taxa, and then fitted models with taxonomic ranks from phylum to genus, thus including 893 taxa.

To reduce noise and dimensionality, a taxa-aware FS step was performed prior to fitting predictive models. The method is developed in the following section 4.2.2.

Taxa-aware feature selection

Due to the hierarchical structure of taxonomy, redundancy occurs when including several taxonomic ranks in analyses [116]. Prior knowledge of which ranks are the most relevant for inclusion is often limited, leading to high-dimensional datasets. Therefore, taking into account taxonomic hierarchy during FS can help further reduce dimensionality by limiting selected taxa to their most relevant ranks [172, 5, 99].

I modified the guided regularized random forest (gRRF) FS algorithm [48] to consider the taxonomic structure. To do this, I added a term reflecting the importance of taxa phylogenetically related to the focal taxon i when calculating its regularization term λ_i. The original λ_i,

$$\lambda_i := 1 - \gamma + \gamma \frac{Imp_i}{Imp*},\ \gamma \in [0,1], \tag{4.1}$$

was hence defined as

$$\lambda_i := 1 - \gamma + \gamma \Big(\frac{Imp_i}{Imp*}\Big)^{1-k} \Big(\frac{Imp_i}{max(Imp_j \,|\, j \in b)}\Big)^{k},\ k \in [0,1], \tag{4.2}$$

with b the subset of variables in the same taxonomic branch as variable i. For variables not describing a taxon, e.g., a metadata, λ_i was calculated as in Equation 4.1.

b is defined relatively to i to comprise all directly coarser and finer taxa. If i was the finest taxonomic level included for FS, sister levels of i were added to b. For instance, if the family, genus and species levels were used, b was defined for each level as:

- family: the family and all its genera;
- genus: the genus, the family it belongs to and all its species;

- species: the genus it belongs to and all species of that genus.

Both γ and k were tuned to evaluate how much weight should be given to gRRF Gini importances and to the taxonomic term in Equation (4.2), respectively. For each model sequence, 121 combinations of FS parameters were tested.

Model interpretation with endoR

The final fitted model was processed with endoR: variables were discretized in $K = 2$ categories, bootstrapping was performed on $B = 100$ resamples, and α was chosen to maximize the number of predicted samples while being as small as possible (Figure 4.4 A).

Samples were classified into four groups using k-means according to their affiliation to decisions' sample support and response variable (clustering repeated a maximum of 500 times to reach stable groups). The lowest number of clusters that minimized within-group variances was chosen (Figure 4.2).

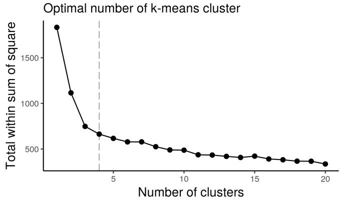

Figure 4.2: Variance of k-means clusters of samples, built based on the presence of *Methanobacteriaceae* and predictions of the stable decision ensemble. Dashed line: number of k-means clusters selected.

4.3 Results

4.3.1 Gut microbial diversity maps onto the enterotype landscape

Metagenomes gathered for the following analysis had been collected for 26 studies from 2203 individuals worldwide, living in 23 countries in total (Appendix B, Table B.2). Participants were aged from 19 to 84 years old, with a median and mean age of 33 and 40 years old, respectively (no age information was reported for 528 individuals), and with a BMI ranging from 16.02 to 36.41 kg.m^{-2}, median and mean BMI being 23.27 and 24.03 kg.m^{-2}, respectively. Finally, 76.53 % of sampled individuals were from westernized populations, in the sense of living an urban lifestyle with a diet comprising fewer carbohydrates and more fat, sugar and animal products compared to rural populations [183, 42].

I first explored the spread of samples along the enterotype landscape [10, 38]. Similar to previous findings [10, 38], the Jensen–Shannon distance calculated from the relative abundances of genera separated observations according to gradients of enrichment in *Bacteroides* and *Prevotella* (Figure 4.3 A-B). However, samples did not strongly cluster, as shown by the within-group silhouette scores below 0.5, indicating weak clustering [38, 119] (Figure 4.3 D-G). This was to be expected due to the heterogeneity of studies included in the meta-analysis and is consistent with the low silhouette scores reported for these same data [176]. Clustering in three groups resulted in sample groups consistent with the ETB, ETF, and ETP enterotypes previously reported as mapping onto the gradients in *Bacteroides* and *Prevotella* relative abundances [10, 38] (Figure 4.3 A-C). Since the ETF enterotype has been positively associated with higher relative abundances of *M. smithii* [38], despite the enterotypes low homogeneity, they were included in further analysis to verify their association with the methanogen.

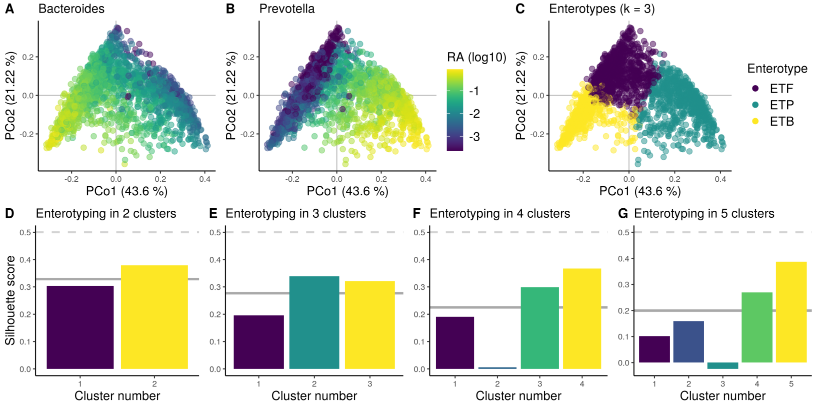

Figure 4.3: Enterotyping of data. A-C/ Principal coordinate analysis ordination of the Jensen-Shannon distance matrix used to cluster samples. A-B/ Samples colored by relative abundance (RA) of *Bacteroides* and *Prevotella*, respectively. For each genus, a pseudocount equal to the minimal non-null RA was given to samples for which the genus was not detected (i.e., RA = 0) to calculate the log. C/ Samples colored by enterotype cluster [10, 38]; ETF: *Firmicutes*, ETB: *Bacteroides*, ETP: *Prevotella*; colors correspond to those on E. D-G/ Average silhouette score within each cluster (bar) and across clusters (thick line). Dashed line: threshold above which clustering strength is moderate.

4.3.2 The occurrence of *Methanobacteriaceae* is not associated with age, BMI or gender

To determine whether gender, age or BMI are associated with methanogens in human guts, I fitted RF models on the 748 samples with complete metadata information. On average, the best model Cohen's κ (0.56±0.05) was lower than for models fitted on all 2203 samples that did not include age and BMI in the set of predictors (0.60±0.02, respectively, Appendix B, Table B.3). As this could be due to the fewer observations available for model training, I sought to determine whether gender, age and BMI were important for predictions. Across all 10 CV repetitions of the best predicting model fitted on the 748 samples, none of these variables was ever selected. Therefore, gender, age, and BMI were determined as not important to predict methanogens' presence from this set of observations and were excluded for further analyses.

4.3.3 Members of the *Oscillospiraceae* and CAG-138 families determine the occurrence of *Methanobacteriaceae* in human guts

I trained several tree ensemble models to predict the presence of *Methanobacteriaceae* in human guts using taxonomic and metabolic pathways relative abundances, available metadata on individuals, and enterotypes (metadata are listed in Appendix B, Table B.1). The selected model had an expected accuracy of 0.83±0.01 and Cohen's κ of 0.61±0.02 on unseen observations (see details on model training and selection in section 4.2.2 and Appendix B, Table B.3). The final model, fitted on all observations, resulted in 75 features selected by the taxa-aware gRRF algorithm, to which metadata were added to fit the predictive RF.

A stable decision ensemble was extracted from the predictive model using endoR with $\alpha = 15$ on $B = 100$ bootstrap resamples. It comprised 23 decisions that could make predictions on 2057 samples, out of the total 2203 (Figure 4.4 A), with an average error of 0.32±0.08 and support of 0.27±0.07 (Figure 4.4 B).

A total of sixteen features were used in decisions to predict the presence of *Methanobacteriaceae*. The Phil-1 genus (family CAG-138, order *Christensenellales*) had the highest importance, and the CAG-138 family, to which the genus belongs to, was ranked shortly after in terms of feature importance (Figure 4.5 A). The glycolysis IV pathway was the only metabolic pathway present in the decision ensemble; it had the second highest importance and was negatively associated with the

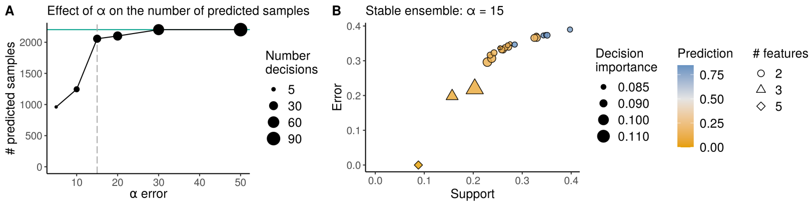

Figure 4.4: Processing of the final model with endoR: choice of α and decision characteristics. A/ α was selected to maximize the number of samples that could be predicted with the decision ensemble, while minimizing the number of decisions in the ensemble; therefore, $\alpha = 15$ was used to obtain the stable decision ensemble (grey dashed line). Green line: $n = 2203$, the total number of samples in the dataset. B/ Characteristics of decisions in the stable decision ensemble. Shape: number (#) of features used in decisions. Color: predictions < 0.5 correspond to the absence of methanogens.

presence of methanogens (Figure 4.5 A). Finally, the *Oscillospiraceae* family was over-represented, with four taxa of this family used in decisions, and together with the CAG-382 family, they accounted for five members of the *Oscillospirales* (Figure 4.5 A). No metadata was used in decisions to predict samples, meaning that microbial features were sufficient to discriminate samples where *Methanobacteriaceae* were detected from those where they were not.

The *Holdemanella* genus (order *Erysipelotrichales*, class *Bacilli*, phylum *Firmicutes*) was the most connected node in the interaction network, sharing edges with eight distinct features. Nonetheless, the glycolysis IV pathway and Phil-1 genus both had a node degree of 7, while also having the highest feature importances and total interaction importance, i.e., the sum of interaction importances to which they participate. The CAG-170 sp002404795 (family *Oscillospirales*) was the next feature with both highest feature importance and degree. Altogether, this suggests that bacterial markers of the presence of *Methanobacteriaceae* are important due to their interactions with other variables.

The glycolysis IV pathway was negatively associated with *Methanobacteriaceae* and used in decisions such that low relative abundances of the metabolic pathway, together with high relative abundances of *Oscillospirales* or CAG-138 were predictive of the presence of methanogens in samples.

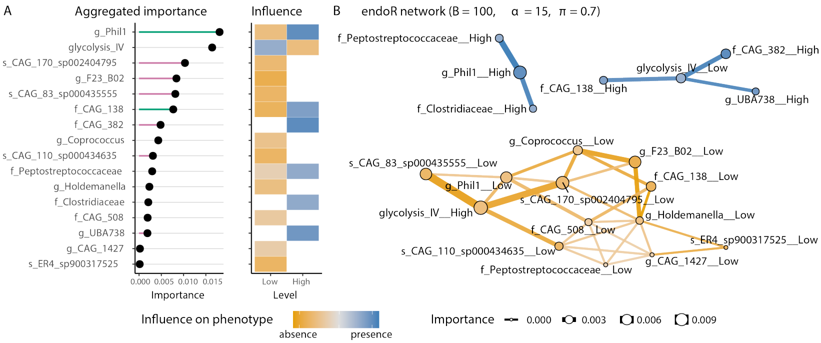

Figure 4.5: Feature importances and interaction network predictive of *Methanobacteriaceae*'s presence. A/ Global feature importance and influence per level. Pink bars: taxa from the *Oscillospiraceae* family or *Oscillospirales* order. Green bars: taxa from the CAG-138 family, order *Christensenellales*. B/ Interaction network. The color legend is common to A and B.

4.3.4 Humans gut microbiomes are positioned on a gradient favorable to colonization by *Methanobacteriaceae*

Samples were clustered into four groups using k-means. Clustering was performed based on the detection of *Methanobacteriaceae* in metagenomes and predictions of the decision ensemble (Figure 4.6 and Figure 4.2).

The presence of *Methanobacteriaceae* was associated with high relative abundances of members of the CAG-138 family (Figure 4.6, group A). The majority of samples also had high relative abundances *Peptrostreptococcaceae* or *Clostridiaceae*, together with the higher relative abundances of members of the CAG-138 family. In addition, a subset of samples also showed lower relative abundances of the glycolysis IV metabolic pathway accompanied by higher relative abundances of members of the *Oscillospirales* order, notably of the *Oscillospiraceae* family, and CAG-138 family. Together, these observations suggest a strong association between *Methanobacteriaceae* and the CAG-138 family. They also suggest an association between *Oscillospirales* and *Methanobacteriaceae* when the relative abundance of the glycolysis IV metabolic pathway is low. Samples in group A were mainly from the ETF enterotype, with 51 % of samples belonging to this enterotype and only 14 % to the ETB enterotype. This was significantly divergent from the proportions in the whole dataset where 28, 32, and 40 % of samples were from the ETB, ETP, and ETF enterotypes respectively (χ^2-test, $\chi^2 = 434.1$, df = 6 and p-value $< 2.2 \cdot 10^{-16}$).

In contrast, a group of samples where *Methanobacteriaceae* were not detected (98 % absent), were characterized by lower abundances of all important taxa and, to a certain extent, higher relative abundances of the glycolysis IV metabolic pathway (Figure 4.6, group D). This group mostly comprised samples of the ETB enterotype (60 % ETB versus 23 % ETF and 14 % ETP), and a majority of ETB samples belonged to this group (54 % of them). Furthermore, westernized individuals were over-represented in group D, accounting for 96 % of samples.

The k-mean algorithm also distinguished a second group of samples where *Methanobacteriaceae* were not detected (Figure 4.6, group C, 97 % of samples without *Methanobacteriaceae*). Similarly, this group was characterized by lower relative abundances of the most important taxa and higher relative abundances of the glycolysis IV pathway. However, the *Holdemanella* genus was not specifically at lower abundances as in group D. Furthermore, both the enterotype and westernized distributions were reflective of those in the whole dataset (25 % ETB, 34 % ETF, and 41 % ETP in group C versus 28, 32, and 40 % respectively in the dataset; and 73 % westernized samples in group C versus 77 % in the population).

A last group of samples where methanogens had not been detected was heterogeneous, presenting either none of or a mix of the aforementioned relative abundance patterns (Figure 4.6, group B). Despite the non-detection of methanogens in samples of group B, 235 of them presented the same patterns as samples in group A, with higher relative abundances of members of the CAG-138 family and *Oscillospirales* order. Samples that displayed mixed patterns were disparately enriched in certain important taxa and depleted in others.

Finally, no prediction could be made on samples that were not part of any decision support (in total 146 samples could not be predicted by the stable decision ensemble, 121 without methanogens belonging to group B and 35 with methanogens belonging to group A). These samples were mostly from the ETP enterotype (60 and 52 % of samples where *Methanobacteriaceae* were detected or not, respectively, detected were ETP), and those with methanogens comprised more ETF than ETB enterotypes (respectively, 26 and 14 %), while samples without methanogens comprised slightly more ETB than ETF (respectively, 28 and 20 %).

4.4 Discussion

In this chapter, I investigated a large dataset of gut microbiomes sampled worldwide to identify patterns of bacterial markers associated

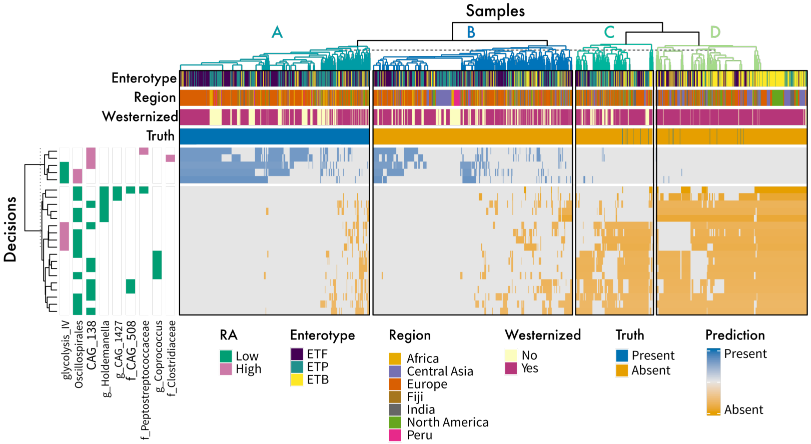

Figure 4.6: Clustering of samples according to predictions of the presence of *Methanobacteriaceae* by the stable decision ensemble. Samples were clustered in four groups using k-means based on the presence of methanogens (Truth) and predictions by the stable decision ensemble (heatmap). Groups are labelled A-D. Decisions are displayed by a matrix (left), where each feature used is colored by its relative abundance (RA), e.g., the bottom decision is low RA of CAG-138 and *Oscillospirales*. Members of the *Oscillospirales* order and CAG-138 family are grouped for concision.

with methanogens. Metagenomes were used to predict the presence of *Methanobacteriaceae* with relative abundances of bacteria, described by several taxonomic levels to consider specialized and general interactions, and metabolic pathways. Information on sampled individuals and original studies part of the meta-analysis were also included to correct for covariates.

A taxa-aware FS step showed that methanogens' presence is not associated with either age, gender, or BMI in the broad population studied. Previous reports of correlations between *M. smithii*, and age [233, 28, 245, 154] or BMI [144, 9, 157, 80, 201, 106, 28, 150, 249, 228] may be population-specific, or false positives due to confounding factors or utilization of inadequate methods. However, with reference to the association between age and methanogens, it must be noted that only adults and elders were included in the present analysis, preventing possible comparison with other stages of life, i.e., infants, children, and teenagers.

Processing of the final predictive model with endoR highlighted the importance of combinations of certain bacterial families and of the glycolysis IV pathway to predict the presence of *Methanobacteriaceae*.

In particular, my results showed that the Phil-1 genus and its family, the CAG-138, *Christensenellales* order, were the most important taxa associated with methanogens. The *Christensenellaceae* family, order *Christensenellales*, was the first family of the order to be described and, to date, is the only family to comprise isolates. It has also been repeatedly associated with *M. smithii* [80, 94, 230, 114]. Therefore, my findings suggest that at a broader population scale, the CAG-138 family is more strongly associated with methanogens than *Christensenellaceae* are. Moreover, since CAG-138 relative abundances were not included in early studies, *Christensenellaceae* relative abundances may have been proxies for CAG-138 abundances, hence confounding associations.

Multiple members of the *Oscillospiraceae* family, the CAG-382 and CAG-272 families, order *Oscillospirales*, were used to predict the presence of *Methanobacteriaceae*. An in-depth study of the genome of the co-abundance gene group *CAG-83*, *Oscillospiraceae* family, predicted it to be a glycan-degrading bacterium likely to produce butyrate and a slow grower [83]. Members of the *Oscillospirales* order may thus produce H_2 during fermentation and consequently, be involved in a syntrophic relationship with methanogens in gut microbiomes. In addition, *M. smithii*, the most abundant and prevalent *Methanobacteriaceae* in human guts, is also a slow grower associated with slow transits [121, 238]. Thus, the methanogen may share niche preferences with members of the *Oscillospiraceae* family, which are associated with slow transits due to their long-predicted replication time [232, 83].

Microbial patterns associated with colonization of gut microbiomes by methanogens were identified by visualizing decisions and predictions across samples. Relative abundances of important taxa identified by endoR formed a gradient. At one end, all taxa were at high relative abundances and *Methanobacteriaceae* were detected in all samples, ranging to the other end, where all relative abundances were low and methanogens were detected in none of the samples. Conversely, relative abundances of the glycolysis IV pathway were low or high in samples with or without *Methanobacteriaceae*, respectively. Although not included in decisions and weakly supported by their within-group variance, enterotypes largely followed this gradient. The ETF group was over-represented among samples with *Methanobacteriaceae*, while the ETB enterotype was over-represented among samples without methanogens. Since the vast majority of the identified important taxa belonged to the *Firmicutes* phylum, the enterotype landscape may follow the gut colonization gradient by *Methanobacteriaceae* due to the association of some *Firmicutes* with *Methanobacteriaceae*. Despite the number of studies that reported an enrichment in methanogens in the guts of non-westernized populations [170, 147, 42], my analysis did not confirm these findings. Nonetheless, samples depleted in taxa associated with *Methanobacteriaceae* were found to be predominantly westernized. Notably, across the three groups of samples where no *Methanobacteriaceae* were detected, the group comprising almost only westernized samples was characterized by lower relative abundances of the *Holdemanella* genus, from the *Erysipelotrichales* family and *Bacilli* class, compared with samples of the two other groups.

Finally, results supported the importance of bacterial taxa over metabolic pathways and host characteristics for the colonization of guts by *Methanobacteriaceae*. This indicates that metabolic pathways supporting methanogens' growth may be ubiquitous in human guts and that at a large population scale, host characteristics may not be confounded with microbiome composition. Furthermore, specific interactions between methanogens and identified taxa must occur consistently across individuals worldwide, as shown by individuals of all countries being mixed along the gradient.

A number of samples showed mixed patterns or could not be predicted by the decision ensemble. Additional or alternative factors must influence the colonization of guts by *Methanobacteriaceae* in those individuals. The RF model may not have captured them due to a lower representation of such samples in the dataset or a lack of information, e.g., specific diets. For example, consumption of raw milk by children has been hypothesized to be a source of *M. smithii*, resulting in higher relative abundances of methanogens [231]. Additional host information could help characterize microbial environments favoring the colonization of humans by *Methanobacteriaceae* in the future.

Collectively, these results give evidence for the complex interactions that occur in the human gut, here resulting in varying occurrences of methanogens. They additionally highlight the value of large-scale analyses to disentangle host characteristics from microbiome composition. For instance, several studies have reported associations between *M. smithii* and host BMI, including anorexia [144, 9], leanness [157, 80, 201, 106, 28], and obesity [150, 249, 228]. They additionally suggested that mediating methanogens' relative abundance in humans may alleviate such phenotypes [28, 148]. As shown by my results on a large population, BMI is not associated with *Methanobacteriaceae*. Even though methanogens may be markers of host phenotypes, the bacteria with which they interact are the most likely to affect the host, and methanogen's abundances may have been confounded in the studied populations with bacteria truly associated with BMI. Furthermore, to correctly investigate microbiomes, associations must be identified using appropriate methods, therefore without classical statistical analysis such as Spearman's correlation, and by extracting as much information from analyses as possible. This will better support experimental designs aiming to characterize underlying relationships in a second step. EndoR enables to overcome the low interpretability of tree ensemble models, as here illustrated with predictions of *Methanobacteriaceae*'s presence. This interpretation method can be applied to various problems to enhance our understanding of human gut microbiomes.

Chapter 5

Syntrophy via interspecies H_2-transfer between *Christensenella* spp. and *Methanobrevibacter smithii*

The aim of this chapter is to provide evidence for the biological relevance of my meta-analysis of metagenomes. This will be illustrated through the example of the *M. smithii* - *Christensenellaceae* relationship, which has previously been suggested as a result of sequence data analysis [80, 94, 230, 114], and that is in line with my finding of an association between *M. smithii* and the CAG-138 family, order *Christensenellales*, for which no isolate is available. Microscopy images, and gas and SCFA concentrations acquired during the course of co-culture experiments of the methanogen with members of the *Christensenellaceae* family will be contrasted with results from cocultures of the methanogen with *Bacteroides thetaiotaomicron*, an H_2-producer ubiquitous to human guts but non-associated with *M. smithii*.

Parts of this chapter were originally published in Ruaud and Esquivel-Elizondo, 2020 [194]. The following text was adapted from the original manuscript for this dissertation. I conducted all experiments in collaboration with Dr Sofia Esquivel-Elizondo (i.e., experimental design, making of experiments, data acquisition), data analysis and making of figures was done by me alone. All author contributions relevant to this chapter are detailed in Appendix C, Table C.1.

5.1 Introduction

Several studies have reported a correlation between *M. smithii* and the *Christensenellaceae* family, order *Christensenellales*, in the human gut [80, 94, 230, 114]. These bacteria are of particular interest for humans as they have been consistently associated with leanness, though the underlying mechanisms remain unknown [80, 94, 41, 74, 81]. Moreover, both the bacteria and the methanogen are heritable taxa, meaning that the host genetics explains a small but significant part of their relative abundance in the gut microbiome [80, 81, 229, 130, 19].

Christensenellaceae have only recently been described [159] and no more than five species, all from the *Christensenella* genus, have been isolated to date [159, 163, 164, 127, 132]. Nonetheless, many MAGs have been assembled in the last few years, such that the family now comprises 17 genera referenced in the GTDB genome database [175] (accessed on April 17th, 2021). Given that the cultured representatives of *Christensenellaceae* ferment simple sugars [159, 127] and that their genomes contain hydrogenases [192], it is likely that they produce H_2. Their association with *M. smithii* could thus be due to the utilization by the archaea of the bacterial H_2 as a substrate for methanogenesis.

Understanding how the methanogen interacts with members of the *Christensenellaceae* family would not only provide support to sequence-based findings, but would also provide insight into the mechanisms by which *Methanobacteriaceae* are associated with host phenotypes.

I explored the association between *Christensenella* spp. and

M. smithii via co-culture experiments. To accomplish this, i) gas production and consumption, as well as SCFA production, were compared between mono- and co-cultures to estimate gas flows and bacterial fermentation changes, and ii) imaging via confocal and scanning electron (SEM) microscopy was performed to assess physical interaction between microorganisms. Due to the culture fastidousness of *M. smithii*, experiments focused on *Christensenella minuta*, the most abundant cultured *Christensenella* in human guts. Experiments were then extended to *Christensenella timonensis* and *Christensenella massiliensis*. Moreover, the strength of the association was assessed by comparing co-cultures of *M. smithii* and *C. minuta* with co-cultures of *M. smithii* and *Bacteroides thetaiotaomicron*. The methanogen can grow in the laboratory from the H_2 provided by *B. thetaiotaomicron* [227, 113, 169], a common gut commensal never found associated with *M. smithii* in the human gut. This bacterium is thus a good control to compare how non-associated versus associated bacteria support the archaeon in co-cultures. Results showed that *Christensenella* spp. outperform *B. thetaiotaomicron* in supporting the growth of *M. smithii* via interspecies H_2-transfer. In addition, *M. smithii* directed the metabolic output of *Christensenella* spp. towards less butyrate and more acetate. In summary, this work demonstrates that the association between *Christensenellaceae* and *M. smithii*, repeatedly detected from metagenome data analysis, is most likely due to efficient H_2-transfer favoring gut colonization by the methanogen.

5.2 Methods

5.2.1 Culturing of methanogens and bacteria

M. smithii DSM-861, *C. minuta* DSM-22607, *C. massiliensis* DSM-102344, *C. timonensis* DSM-102800, and *B. thetaiotaomicron* VPI-

5482 were obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ; Braunschweig, Germany). Each culture was thawed and inoculated into Brain Heart Infusion (BHI) medium (Carl Roth, Karlsruhe, Germany) supplemented with yeast extract (5 g/L), reduced with L-Cysteine-HCl (0.5 g/L) and Ti-NTA III (0.3 mM), and buffered with sodium bicarbonate (42 mM, pH 7, adjusted with HCl 6 M). 10 mL cultures were grown at 37 °C without shaking in Balch tubes (total volume of 28 mL) under a headspace of N_2:CO_2 (80:20 % v/v) in the case of the bacteria, and H_2:CO_2 (80:20 % v/v, pressure adjusted to 2 bar) for *M. smithii.* When initial cultures reached exponential growth, and before floc formation, they were transferred into fresh medium and these transfers were used as inocula for the experiments described below.

5.2.2 Co-culture conditions

M. smithii was co-cultured with *C. minuta*, *B. thetaiotaomicron*, *C. massiliensis*, or *C. timonensis*, and in parallel, each microorganism was grown in mono-culture (Table 5.1). Prior to inoculation, one-day old cultures of bacterial species, or four-day old cultures of *M. smithii*, were adjusted to an OD600 of 0.01 with sterile medium. For the co-cultures, 0.5 mL of each adjusted culture were inoculated into 9 mL of fresh medium. For the mono-cultures, 0.5 mL of the adjusted culture and 0.5 mL of sterile medium were combined as inoculum. For negative controls, sterile medium was transferred as a mock inoculum. Headspaces were exchanged with 80:20 % (v/v) of N_2:CO_2 or H_2:CO_2 and pressurized at 2 bar or atmospheric pressure (i.e., 0.98 bar, Table 5.1). Each batch of experiments was carried out once with 3 biological replicates per culture conditions (Table 5.1).

Batch	Pressure (bar)	Headspace (80:20 % v/v)	Culture inocula
1	2.0	N_2:CO_2	*C. minuta* *M. smithii* / *C. minuta* *B. thetaiotaomicron* *M. smithii* / *B. thetaiotaomicron* sterile medium (negative control)
		H_2:CO_2	*M. smithii* sterile medium (negative control)
2	0.98	N_2:CO_2	*C. minuta* *M. smithii* / *C. minuta* sterile medium (negative control)
		H_2:CO_2	*M. smithii* sterile medium (negative control)
3	2.0	H_2:CO_2	*C. minuta* *M. smithii* / *C. minuta* *M. smithii* sterile medium (negative control)
4	0.98	N_2:CO_2	*C. massiliensis* *M. smithii* / *C. massiliensis* *C. timonensis* *C. timonensis* / *M. smithii* sterile medium (negative control)
		H_2:CO_2	*M. smithii* sterile medium (negative control)

Table 5.1: Total pressure, headspace composition, and culture inocula for each batch of experiments.

5.2.3 Imaging

For confocal microscopy, SYBR ® Green I staining was performed as previously described [125] with the modifications detailed in Appendix C, Additional methods. Imaging by confocal microscopy (LSM 780 NLO, Zeiss) was used to detect the autofluorescence emission of coenzyme F420 of *M. smithii* and the emission of SYBR® Green I (Appendix C, Additional methods). Images were acquired with the ZEN Black 2.3 SP1 software and processed with FIJI [199]. Micrographs are representative of all replicate cultures within each experimental batch. The preparation of samples for scanning electron microscopy is described in Appendix C, Additional methods. Cells were examined by Jürgen Berger (Electron microscopy facility, Max Planck Institute for Developmental Biology) with a field emission scanning electron microscope (Regulus 8230, Hitachi High Technologies, Tokyo, JPN) at an accelerating voltage of 10 kV.

5.2.4 Gas and SCFA measurements

Headspace concentrations of H_2, CO_2, and CH_4 were measured with a gas chromatograph (GC) (SRI 8610C; SRI Instruments, Torrence, USA) equipped with a packed column at 42 °C (0.3-m HaySep-D packed Teflon; Restek, Bellefonte, USA), a thermal conductivity detector (TCD) at 111 °C, and a flame ionization detector. The gas production and consumption were estimated from the total pressure in the vials (ECO_2 manometer; Keller, Jestetten, Germany) and the gas concentrations in the headspace using the ideal gas equation. The concentrations are given in mMol of gas in the headspace per liter of culture.

SCFA measurements were performed with liquid samples (0.5 mL) filtered through 0.2 μm pore size polyvinylidene fluoride filters (Carl Roth, GmbH, Karlsruhe, GER). SCFA concentrations were measured with a CBM-20A high performance liquid chromatography (HPLC) system equipped with an Aminex HPX- 87P column (300 x 7.8 mM, BioRad, California, USA), maintained at 60 °C, and a refractive index detector. A sulfuric acid solution (5 mM) was used as eluent at a flow rate of 0.6 mL.min^{-1} (40 bar column pressure). Calibration curves for acetate and butyrate were prepared from 1.25 to 50 mM using acetic acid and butyric acid, respectively (Merck KGaA, Darmstadt, Germany). No other fatty acids were detected (Appendix C, Additional methods). The SCFA concentrations were estimated with the Shimadzu LabSolutions software.

5.2.5 Statistical analyses

I used Wilcoxon rank sum tests to compare gas production between cultures after 6 days of growth. When more than one culture condition (i.e., headspace composition and pressure,, in mono- or co-culture, Table 5.1) was included in the comparison, I instead performed an ANOVA followed by Tukey's post-hoc test to discriminate between the effects of the different conditions.

SCFA concentrations were compared using a two-way ANOVA such that the culture conditions (i.e., headspace composition and pressure, Table 5.1) and the sample (mono- or co-culture) were evaluated to explain the variance of butyrate and acetate concentrations after 6 days of growth. *p*-values were adjusted using the Benjamini-Hochberg method. A Tukey's post-hoc test was performed to discriminate between the effects of the different conditions. All statistical analyses were done in R using the stats R package.

5.2.6 Comparison of expected (theoretical) versus measured CH_4 production in co-cultures

We used the stoichiometry of hydrogenotrophic methanogenesis: $CO_2+4H_2=CH_4+2H_2O$, to calculate the amount of CH_4 that could be produced from the estimated amount of H_2 consumed in each sample (Table 5.2). For this, we used the mono-cultures of bacteria as references and assumed H_2 production in co-culture was equivalent to that in mono-culture. We estimated the H_2 consumed after 6 days for each replicate as the difference between the averaged H_2 concentrations in mono-cultures and the concentration measured in co-culture (i.e., unconsumed H_2). The estimated H_2 consumed was then divided by 4 in order to obtain the theoretical amount of CH_4 that could have been produced via hydrogenotrophic methanogenesis.

5.3 Results

5.3.1 *M. smithii* consumes the H_2 produced by *C. minuta*

To validate our hypothesis that *Christensenellaceae* produce H_2 which *M. smithii* can grow on, we tracked gas concentrations in mono- and co-cultures of the methanogen with *C. minuta*, the most abundant *Christensenellaceae* in the human gut, and *B. thetaiotaomicron*, a commensal H_2-producer of the human gut [227, 113, 169].

M. smithii did not grow in mono-culture when no H_2 was supplied (80:20 % v/v N_2:CO_2 headspace, Figure 5.1 b), but did when it was provided in excess (i.e., 80:20 % v/v H_2:CO_2 atmosphere at 2 bars). After 6 days, the methanogen had produced on average 9.0 ± 1.0 mmol.L^{-1} of CH_4 (Figure 5.1 b and Figure C.1 b). *C. minuta* had produced on average 7 times more H_2 than *B. thetaiotaomicron* in mono-culture (after 6 days, H_2 concentrations were 14.2 ± 1.6 mmol.L^{-1} versus 2.0 ± 0.0 mmol.L^{-1}, Figure 5.1 a and d and Figure C.1 a; Wilcoxon rank sum test, p-value = 0.1). Accordingly, *M. smithii* in co-culture with *C. minuta* outgrew the co-culture with *B. thetaiotaomicron* (respectively 5.8 ± 0.5 mmol.L^{-1} and 1.1 ± 0.0 mmol.L^{-1} of CH_4 measured; Wilcoxon rank sum test, p-value = 0.1; Figure 5.1 c and e and Figure C.1 b). The methanogen consumed most H_2 from bacterial fermentation, as shown by the very low H_2 concentrations in co-culture (on average across all time points and replicates, H_2 concentrations were 0.5 ± 0.6 mmol.L^{-1} in co-cultures with *C. minuta*, and 0.1 ± 0.1 mmol.L^{-1} with *B. thetaiotaomicron*; Figure 5.1 c and e and Figure C.1 a).

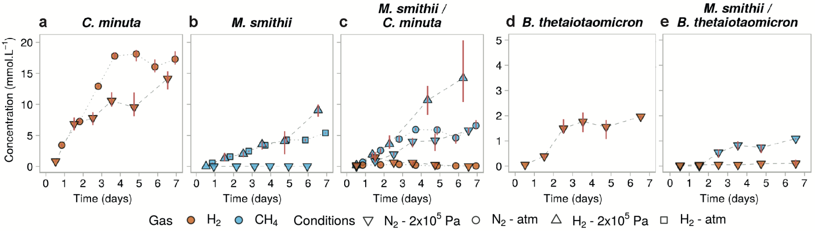

Figure 5.1: Gas concentrations over time in mono- and co-cultures of *M. smithii*, *C. minuta*, and *B. thetaiotaomicron* grown under different conditions. Average of the 3 biological replicates for each condition (points), and minimal and maximal values (red bars). In conditions where H_2 was provided in excess (H_2 - 2 bar and H_2 - atm, headspace initially composed of 80:20 % H_2:CO_2), its concentrations are not shown for scale reasons. Initial concentrations of H_2 in conditions where it was not provided in the headspace were undetectable (N_2 - 2 bar and N_2 - atm, headspace initially composed of 80:20 % N_2:CO_2) and stayed null in the mono-cultures of *M. smithii* (not shown). CH_4 concentrations in the bacterial mono-cultures were undetectable and are not shown as well. Panels a-c share the same y-scale, as do panels d-e.

Gas solubility increases with pressure as described by Henry's law. Consequently, gas consuming microorganisms are predicted to grow better in a pressurized environment [55, 12]. We tested whether *C. minuta* would similarly support the growth of *M. smithii* even at lower pressure, i.e., starting at atmospheric pressure (0.98 bar) instead of 2 bar, the recommended pressure to grow *M. smithii* [12, 93, 55]. H_2 production was higher in mono-cultures of *C. minuta* at atmospheric pressure compared to 2 bar (respectively 17.28±1.12 and 14.15±1.56), supporting *M. smithii*'s growth to a similar extent in coculture (ANOVA followed by Tukey's post-hoc test, adjusted *p*-value = 1.0; Figure 5.1 c, Figure C.1 b).

5.3.2 *M. smithii* colonizes flocs formed by *C. minuta*

A striking phenotype observed from cultures containing *C. minuta* was the formation of a biofilm visible with the naked eye from 3 days of growth (Figure 5.2 a and g). We therefore imaged cultures with confocal microscopy and SEM between 3 to 7 days after inoculation (Figure 5.2 and Figure 5.3).

C. minuta's flocs were colonized by *M. smithii* at least as early as 3 days post-inoculation (Figure 5.3 a,b and d, and Figure 5.2 a-c and g-j). *M. smithii* did not aggregate in mono-culture before 7 to 10 days (data not shown). In comparison, *B. thetaiotaomicron* did not flocculate when grown alone (Figure 5.3 c) and displayed very limited aggregation when co-cultured with *M. smithii* (Figure 5.3 e, Figure 5.2 k-n).

As Fick's law of diffusion states that the flux of a metabolite between two microorganisms is directly proportional to the concentration gradient and inversely proportional to the distance [212, 208], we hypothesized that the aggregation of *M. smithii* and *C. minuta* facilitates H_2-transfer between the methanogen and the bacterium. We hence examined whether *M. smithii* joins *C. minuta*'s flocs if provided with H_2 in excess in the headspace at high pressure (i.e., 80:20 % v/v H_2:CO_2 atmosphere at 2 bars). In such conditions, the methanogen does not depend on the bacterium as an H_2 source. *M. smithii* aggregated with *C. minuta* (Figure 5.3 f-g) and CH_4 production was even higher than in

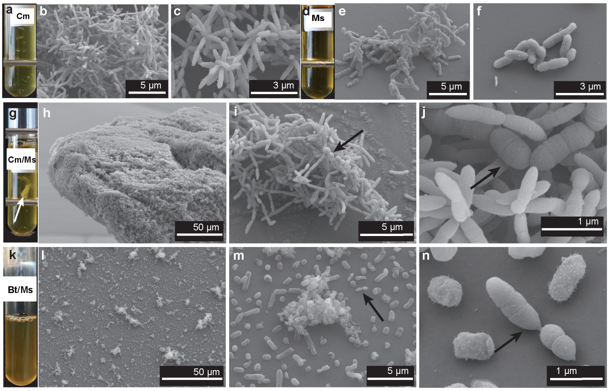

Figure 5.2: Scanning electron micrographs of the cultures at 3-7 days of growth. a, d, g and k: Representative Balch tubes of cultures of *C. minuta* (Cm), *M. smithii* (Ms), *C. minuta* and *M. smithii* (Cm/Ms), and *B. thetaiotaomicron* and *M. smithii* (Bt/Ms) after 7 days of growth. In panel g, the floc formed by Cm/Ms is indicated with an arrow; b-c: Scanning electron micrographs (SEMs) of mono-cultures of *C. minuta* at 5 days of growth; e-f: SEMs of mono-cultures *M. smithii* at 5 days of growth; h-j: SEMs of co-cultures of *C. minuta* and *M. smithii* at 7, 5 and 2 days of growth respectively; l-n: SEMs of co-cultures of *B. thetaiotaomicron* and *M. smithii* at 7 days of growth. Arrows indicate *M. smithii* cells. Metal bars on a, d and j are from the tube rack.

mono-culture under the same condition, reaching 14.2 ± 5.3 mmol.L^{-1} in co-culture versus 9.0 ± 1.0 mmol.L^{-1} in mono-culture after 6 days (ANOVA followed by Tukey's post-hoc test, adjusted p-value = 0.1, Figure 5.1 b and c). This indicates that interspecies H_2-transfer occurs even when H_2 is added to the headspace and that it boosts methanogenesis rather than solely uptaking H_2 provided in the headspace.

Note *C. minuta* and *M. smithii* also aggregated at atmospheric pressure (Figure 5.3 h-i).

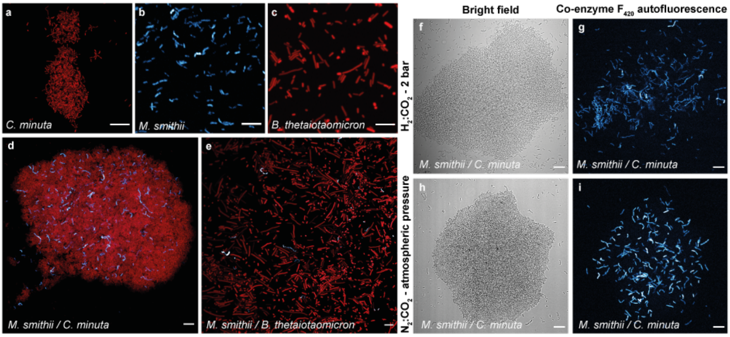

Figure 5.3: Confocal micrographs of mono- and co-cultures of *M. smithii*, *C. minuta*, and *B. thetaiotaomicron* at 3 days of growth. a-e: cultures from Batch 1 (Table 5.1); SYBR® Green I fluorescence (DNA staining) is shown in red and *M. smithii*'s coenzyme F420 autofluorescence is shown in blue; based on gases production, at 3 days of growth, *B. thetaiotaomicron* was already at stationary phase (explaining the elongated cells), *C. minuta* was at the end of the exponential phase and *M. smithii* was still in exponential phase.
f-i: co-cultures from Batches 2 and 3 (Table 5.1) with culture condition indicated on the left.
Scale bars represent 10 μm.

5.3.3 *M. smithii* influences the SCFA production of *C. minuta*

We assessed whether bacterial fermentation was modified in co-cultures due to *M. smithii*. To achieve this, we compared the SCFA concentrations between *C. minuta*'s mono-cultures and co-cultures with the methanogen (i.e., cultures at 2 bar or atmospheric pressure with an 80:20 % v/v N_2:CO_2 or H_2:CO_2 headspace, Table 5.1). Under all tested conditions, acetate and butyrate were the only SCFAs produced by *C. minuta* (among 10 short and medium chain fatty acids screened, Appendix C, Additional methods) and we thus focused our analyses on them.

Butyrate was consistently measured at lower concentrations in co-cultures compared to mono-cultures, with an average difference of concentrations of 1.1 ± 0.24 mmol.L^{-1} after 6 days (Figure 5.4 a-c, Figure C.1 c and Table 5.2; ANOVA, F-value(1) = 161.461 and adjusted *p*-value = 7.7x10^{-8}). The interaction factor between the mono/co-culture conditions and the growth condition was not significantly correlated to butyrate concentrations (ANOVA, F-value(2) = 0.862, adjusted *p*-value = 0.4). Therefore, butyrate production was inhibited in co-cultures regardless of pressure and headspace composition, meaning that the methanogen's presence steadily inhibited *C. minuta*'s fermentation to butyrate.

In addition, acetate production slightly but significantly increased in co-cultures compared to mono-cultures (Figure 5.4 d-f and Figure C.1 d; ANOVA, F-value(1) = 317.41 and adjusted *p*-value = 3.2x10^{-9}). After 6 days, differences in acetate concentrations ranged from +0.7 mmol.L^{-1} at 2 bar under an H_2:CO_2 (80:20 % v/v) atmosphere to +2.2 mmol.L^{-1} at atmospheric pressure under an N_2:CO_2 (80:20 % v/v) atmosphere. These differences significantly varied with the headspace and pressure conditions (the interaction term between the mono/co-culture and the growth condition was significantly correlated to acetate production; ANOVA, F-value(2) = 29.09 and adjusted *p*-value = 3.0x10^{-5}). The effect of *M. smithii* on acetate production was thus larger at lower pressure, when H_2 solubility is predicted to be lower and therefore the gas will be present at lower concentrations in the liquid growth media.

5.3.4 *M. smithii* produces more CH_4 than predicted in theory in co-culture with *C. minuta*

We observed more CH_4 than expected in co-cultures of *M. smithii* with *C. minuta* (Figure 5.1 a-c). The expected quantity of CH_4 in co-culture was calculated by assuming equal H_2 production from *C. minuta* in

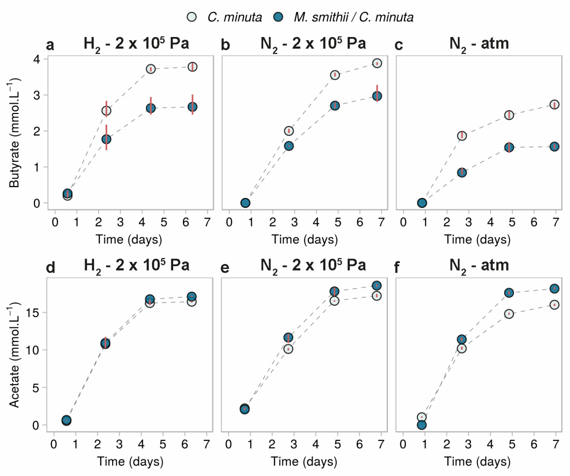

Figure 5.4: SCFA concentrations over time in mono- and co-cultures of *C. minuta* and *M. smithii* grown under different conditions. Short chain fatty acids over time in cultures from batches 1-3 (see Table 5.1). a-c: butyrate concentrations; d-f: acetate concentrations. Only these SCFA were detected among the fatty acids tested (fatty acids from C1 to C8, iso-valerate and iso-butyrate). Points represent the average of the 3 biological replicates for each condition, and red bars join the minimal and maximal values. Mono-cultures of *M. smithii* are not shown as they did not differ from the blanks (negative controls).

both mono- and co-cultures (Table 5.2). This suggests that the bacterium outproduced methanogenesis substrates in the presence of the methanogen. Since the production of acetate yields more H_2 than butyrate production [196, 138], the additional CH_4 observed could originate from the shift in metabolism from butyrate to acetate production by *C. minuta* in co-culture, which would have led to higher H_2 yields for the methanogen.

Condition[a]	H_2:CO_2[b]- 2 bar		N_2:CO_2 - atm		N_2:CO_2 - 2 bar	
	Average	SD	Average	SD	Average	SD
H_2 produced in mono-culture	18.71	9.71	17.28	1.12	14.15	1.56
H_2 not consumed in co-culture[c]	-21.81	0.87	0.08	0.01	0.03	0.00
CH_4 expected in co-culture based on H_2 produced in mono-culture[d]	10.13	0.22	4.30	0.00	3.53	0.00
CH_4 observed in co-culture	14.21	5.33	6.57	0.77	5.81	0.45
Difference between observed and theoretical CH_4	4.08	5.50	2.27	0.77	2.28	0.45
Butyrate difference between co- and mono-culture	-1.11	0.30	-1.21	0.04	-0.91	0.27
Acetate difference between co- and mono-culture	0.68	0.10	2.20	0.22	1.36	0.17

Table 5.2: Analysis of the origin of the high CH_4 produced in co-cultures based on the changes in metabolism of *C. minuta*.

[a] For each condition, the average concentrations and standard deviations (SD) among triplicates after 6 days of growth are given in mmol.L^{-1}.

[b] For the experiments grown under an H_2:CO_2 (80:20 % v/v) atmosphere, the average H_2 concentration measured in the negative controls after 6 days (same gas volume sampled as many times as the inoculated tubes) was subtracted from the concentration measured in the cultures.

[c] Average of the concentration of H_2 in co-cultures to which the average of H_2 concentration in mono-culture of *C. minuta* was subtracted.

[d] This amount is calculated based on the stoichiometry of the hydrogenotrophic methanogenesis reaction: $4\ H_2 + CO_2 = CH_4 + 2$ H2O.

5.3.5 *C. massiliensis* and *C. timonensis*, other members of the *Christensenellaceae* family, support the growth of *M. smithii*

We further investigated the association between *Christensenellaceae* and *M. smithii* by replicating our co-culture experiments with *Christensenella massiliensis* and *Christensenella timonensis* at atmospheric pressure. *C. timonensis* had a slower metabolism such that cultures would reach stationary phase after 7 days of growth (Figure 5.5 d, g, i). Nonetheless, for consistency with experiments with *C. minuta* and *B. thetaiotaomicron*, we stopped the experiments after 7 days (Appendix C, Figure C.1).

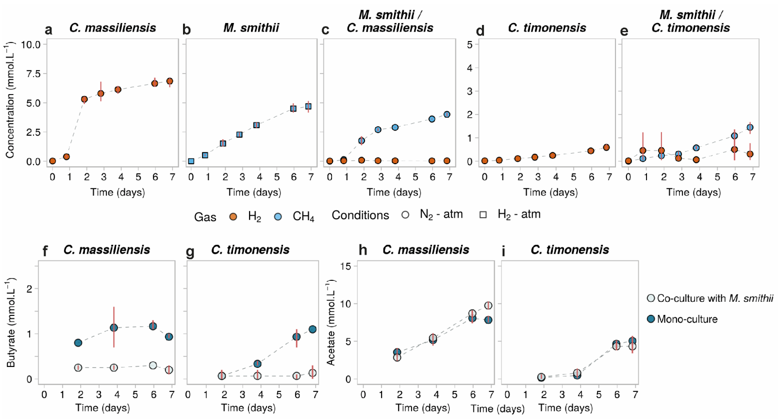

Figure 5.5: Gas and SCFA concentrations in mono- and co-cultures of *C. massiliensis* and *C. timonensis* with *M. smithii*. a-e: H_2 (orange) and CH_4 (blue) concentrations in the headspace in cultures from batch 4 (see Table 5.1); f-g: butyrate and h-i: acetate concentrations in these cultures. Points represent the average of 3 biological replicates, and red bars join the minimal and maximal values. In the mono-cultures of *M. smithii* (b) where H_2 was provided in excess (condition H_2 - atm, headspace initially composed of 80:20 % H_2:CO_2), its concentrations are not shown for scale reasons.

C. massiliensis and *C. timonensis* produced H_2 in smaller quantities than *C. minuta* (concentrations in mono-cultures after 6 days of growth were 6.9 ± 0.5 mmol.L^{-1} for *C. massiliensis* and 0.6 ± 0.1 mmol.L^{-1} for *C. timonensis*, Figure 5.5 a,d and Figure C.1 a). The H_2 was all consumed by *M. smithii* in co-cultures with *C. massiliensis* and 4.0 ± 0.2 mmol.L^{-1} of CH_4 were accordingly produced (Figure 5.5 c and Figure C.1 b). In co-cultures with *C. timonensis*, the methanogen did not uniformly consumed H_2 across replicates, such that it was all consumed in certain but not all across the course of the experiment (Figure 5.5 e; 1.5 ± 0.3 mmol.L^{-1} of CH_4 had been produced after 6 days, Figure C.1 b). Furthermore, *C. massiliensis* and *C. timonensis* formed smaller flocs than *C. minuta* (even if left to grow for longer periods than 7 days), which *M. smithii* colonized (Figure 5.6).

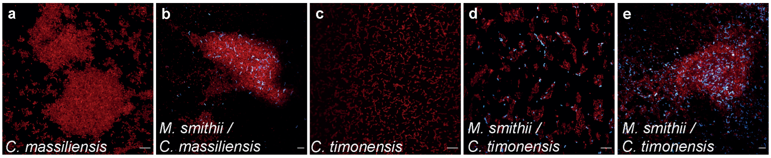

Figure 5.6: Confocal imaging of *C. massiliensis* and *C. timonensis* in mono- and co-cultures with *M. smithii*. Confocal micrographs after 5 days of growth of a: *C. massiliensis*, b: *M. smithii* and *C. massiliensis* in co-culture, c: *C. timonensis*, d-e: *M. smithii* and *C. timonensis* in co-culture. SYBR® Green I fluorescence (DNA staining) is shown in red and, *M. smithii*'s coenzyme F420 autofluorescence is shown in blue. Scale bars represent 10 μm.

Similar to our observations with *C. minuta*, butyrate production was reduced in co-cultures of *C. massiliensis* and *C. timonensis* with *M. smithii* compared with bacterial mono-cultures (Wilcoxon rank sum test, *p*-values = 0.33 for *C. massiliensis* and 0.5 for *C. timonensis*; Figure 5.5 f,g and Figure C.1 c). Butyrate was in fact barely detectable in co-cultures: production dropped from 0.93 ± 0.06 mmol.L^{-1}

and 1.10 ± 0.00 mmol.L^{-1} in mono-cultures of *C. massiliensis* and *C. timonensis* respectively after 6 days, to 0.20 ± 0.14 mmol.L^{-1} and 0.13 ± 0.15 mmol.L^{-1} in co-cultures. Contrasting with the co-cultures of *M. smithii* and *C. minuta*, no significant change in acetate concentrations were observed for *C. massiliensis* and *C. timonensis* (Wilcoxon rank sum tests, *p*-values = 0.2 and 0.8 respectively; Figure 5.5 h,i and Figure C.1 d).

5.4 Discussion

We explored *in vitro* the association previously identified from human gut sequence data between *Christensenellaceae* and *M. smithii. B. thetaiotaomicron* was elected as a control for our experiments as it is used as a H_2 provider to *M. smithii* in labs [113, 227, 169] but has never been reported to correlate with the methanogen in human guts.

Our results showed that *Christensenella* spp. produce H_2 that supports the growth of *M. smithii* in co-cultures. *C. minuta* produced copious amounts of H_2 enabling *M. smithii* to grow as well in co-culture with this bacterium as in mono-cultures with an excess of H_2. Although *C. timonensis* produced less H_2 at atmospheric pressure than *B. thetaiotaomicron* at 2 bar, in co-cultures we measured higher CH_4 concentrations with *C. timonensis* than with *B. thetaiotaomicron.* Furthermore, in co-cultures with *C. minuta*, we observed a greater CH_4 production than expected from bacterial H_2 production in mono-cultures. Finally, *Christensenella* spp. formed flocs in mono-cultures that *M. smithii* colonized in co-cultures. Strikingly, flocs from *C. minuta* were visible with the naked eye. Taken together, these results suggest that *C. minuta*, *C. massiliensis*, and *C. timonensis* outperform *B. thetaiotaomicron* at supporting *M. smithii*'s metabolism via interspecies H_2-transfer. Gas transfer in co-cultures with

Christensenella spp. may be facilitated by close contact within mixed aggregates, to the benefit of the methanogen. In these flocs, the distance between the H_2-producer and consumer is reduced to a minimum, therefore optimizing the flux from one to the other as predicted by Fick's law of diffusion [212, 208].

Consistent with the assumption that *M. smithii* affects bacterial fermentation through H_2 consumption, we observed a change in bacterial SCFA concentrations in co-cultures compared with mono-cultures of *C. minuta*, *C. massiliensis*, and *C. timonensis*. First, butyrate production was inhibited for all three species under all conditions in the presence of *M. smithii.* Second, acetate production was altered to different extents: while it was significantly enhanced for *C. minuta* and mildly increased for *C. massiliensis*, no change was evident for *C. timonensis.* This indicates that the mechanisms by which *M. smithii* influences bacterial fermentation are consistent across *Christensenella* spp. for butyrate production, but are species specific for acetate. The fermentation of distinct bacteria has also been reported to be affected differently by the methanogen in co-cultures [195, 30]. For instance, when the H_2-producing ruminal bacteria *Ruminoccocus albus* and *Ruminococcus flavefaciens* are grown on cellulose with the methanogen and compared with mono-cultures, the fermentation of *R. albus* remains unaltered, while an increase in acetate production occurs for *Ruminococcus flavefaciens.* Our findings related to on *Christensenella* spp.'s metabolism are therefore in line with previous findings.

Based on the theoretical quantities of CH_4 that could have been produced from the H_2 measured in bacterial monno-cultures, unexpectedly high concentrations were measured in co-cultures of *M. smithii* with *C. minuta*. This suggests that *C. minuta* produced additional substrate for *M. smithii*'s methanogenesis. This substrate could be H_2 or formate, or may correspond to direct electron transfer between the two microorganisms. We were unable to confirm boosted H_2 production based on our experiments as the gas was all consumed in co-cultures. Although here I could not detect formate in the supernatant of *C. minuta*'s mono-cultures nor co-cultures, it cannot be excluded as a potential substrate. The genome of the bacterium has been predicted to carry the gene encoding the pyruvate formate lyase, the enzyme that catalyzes the conversion of pyruvate into acetyl-CoA and formate [192], and formate was recently measured in mono-cultures of *C. minuta* grown on glucose [253]. Interspecies formate transfer can take place in co-cultures of H_2-producers and hydrogenotrophic methanogens capable of formate utilization such as *M. smithii* [221, 20], though this is difficult to quantify since formate is rapidly produced and consumed. Moreover, a formate-based symbiosis between *M. smithii* and the ruminal bacterium *Fibrobacter succinogenes* has been reported: in line with our observations in co-cultures of *M. smithii* and *C. minuta*, acetate production was enhanced in the co-cultures of *M. smithii* and *F. succinogenes* [195]. Finally, no evidence of direct electron transfer between *M. smithii* and bacteria has been reported to date. Nonetheless, since this phenomenon can occur in microbial aggregates [208], it is also a potential explanation for the higher metabolism of the methanogen in co-cultures with *C. minuta*.

This work demonstrates that members of the *Christensenellaceae* act as an H_2 source for *M. smithii*, and that this process is enhanced via close physical proximity. Such interactions likely underlie the co-occurrence patterns between the methanogen and *Christensenellaceae* in the human gut microbiome. Further experiments better reflecting the gut environment, e.g., with complex microbial communities or in continuous growth mode (i.e., with continuous influx and removal of growth medium and gases), would allow to deepen our understanding of *M. smithii*'s role in the human gut.

These results support sequence-based analysis studies that have reported these patterns and provide evidence for the biological and practical relevance of results presented in Chapter 4. They also confirm that the methanogen mediates bacterial fermentation of gut microorganisms in various ways, resulting in changes in SCFA that can potentially influence human phenotypes.

Discussion and outlook

The human gut microbiome is a complex environment comprising a myriad of microorganisms interacting together and with their host. Due to culture limitations, insights from the diversity and functioning of microbial components of this environment largely rely on sequence data from stool samples. Among others, they have allowed to characterize archaeal gut diversity [23, 44, 22, 21, 247] and to associate the presence of *Methanobrevibacter smithii*, the most abundant and prevalent human gut archaeon, with host traits such as constipation and slow transit [121, 238], non-western diets [170, 147, 42], and BMI [144, 9, 157, 80, 201, 106, 28, 150, 249, 228]. Since co-culture experiments suggest that *M. smithii* alters bacterial production of short-chain fatty acids (SCFA) [30, 195], and given that these fermentation substrates mediate host metabolism [29, 161, 120, 138, 137, 46], a better understanding of the ecology of gut methanogens is critical to appreciate their impact on humans.

To perform a reliable bioinformatic analysis, one needs the right tools. Comparative studies of gut microbiota, for which associations are inferred between microbes and host traits, lack appropriate workflows enabling accurate and comprehensive analyses. While classical statistical analyses and models produce simple results, they have been shown to be defective in accuracy [236]. Conversely, tree ensemble machine learning models are accurate but generate complex nonintelligible models [225, 115, 116]. A major part of my thesis has been dedicated to develop endoR, a method for interpreting tree ensemble models. Thanks to the use of these models, it produces results more accurate than statistical tests, e.g., Spearman's coefficient of correlation and χ^2 tests, and sparse covariance matrices [70, 124]. Results from endoR are as accurate as those from other tools for interpreting tree ensemble models, e.g., SHAP [140]. However, endoR better scales with high-dimensional data and summarises models into clear figures from which complex interactions can be deduced. It is readily available in R, a statistical language broadly used in microbiome science. In the future, its computation performance could be improved via code optimization in C++, and ideally, the R-package would be translated into a Python package, the second statistical language utilized for biological data analysis.

Tree ensemble models coupled with endoR enabled me exploring associations between human gut methanogens and gut bacterial features in a unique analysis. My results confirmed the strong association between *Methanobacteriaceae* and members of the *Christensenellales* order [80, 94, 230, 114], particularly with the uncultured CAG-138 family. Furthermore, endoR identified multiple associations between methanogens and members of the *Oscillospiraceae*, CAG-382 and CAG-272 families (all from the order *Oscillospirales*). Similar to *M. smithii*, CAG-83, *Oscillospiraceae* family, has been predicted to have a slow replication time and be associated with slow transit [83]. The co-occurrence of methanogens with *Oscillospirales* may thus be due to shared niche preferences, i.e., guts with slow transits so that the washout effect is lower and microorganisms can steadily colonize the environment. Nonetheless, CAG-138 is also predicted to produce butyrate, a SCFA tightly connected to acetate production which produces H_2 [138]. Additional insights into the metabolism of *Oscillospirales* are required to assess the mechanisms underlying their co-occurrence with methanogens in human guts. Moreover, as several species and genera were identified by my analysis, the specific interactions of *M. smithii* with these taxa may be explored to elucidate the adaptation potential of the methanogen to its environment. Finally, an extensive visualization of endoR outputs allowed me to compare results from the model with host traits. I could thus define a gradient of bacterial relative abundances predictive of methanogens' presence in human guts. While samples depleted in all taxa were generally from westernized populations, no clear westernization pattern followed bacterial and methanogen enrichment on the rest of the gradient. Hence, the absence of methanogens in westernized populations can be in part explained by the low relative abundances of *Oscillospirales*, *Christensenellales*, the *Holdemanella* and *Coprococcus*, to cite only a few. In non-westernized populations or westernized individuals not characterized by an ETB enterotype, other factors may prevent the colonization of human guts by *Methanobacteriaceae* when these bacteria are in higher abundances. The given data and analysis could not determine these factors, probably due to a lack of data information, e.g., diet, or sample resolution. Nonetheless, they provide reliable general patterns of bacteria widely co-occurring with *Methanobacteriaceae* that should be investigated with culture-based experiments to associate methanogens with host phenotypes.

I undertook such culture-based experiments to examine the relationship between *M. smithii* and members of the *Christensenellales* order. Among the findings from the meta-analysis was the association between methanogens and the CAG-138 family, order *Christensenellales*, which does not comprise any isolate. Therefore, I conducted experiments with members of the *Christensenellaceae* family, order *Christensenellales*, which have been associated with methanogens in previous studies [80, 94, 230, 114]. This family is of particular interest for humans as it has been repeatedly correlated to leanness [80, 94, 41, 74, 81] and host genetics [80, 81, 229, 130, 19]. At this point it is important to mention that CAG-138 has been described in 2017 [209], hence five years after *Christensenellaceae* [159]. Associations between the later and methanogens, that I could not confirm with my meta-analysis, may thus be due to *Christensenellaceae* serving as proxy to CAG-138 in early analyses, or to their association being true in specific populations only. Nonetheless, my co-culture experiments showed a H_2-based syntrophy between *M. smithii* and members of the *Christensenellaceae* family, particularly strengthened by co-colonization of biofilms. Furthermore, bacterial fermentation was altered in co-cultures and resulted in consistently lower butyrate production, and higher acetate production for one of the three tested species. Altogether, these results support findings from the meta-analysis and provide grounds for characterizing how methanogens may influence host phenotypes through altered bacterial fermentation.

Altogether, my findings complement the current knowledge on interactions between the human gut methanogen *M. smithii* and fermenting bacteria. They support the hypothesis that *M. smithii* preferentially interacts with specific H_2-producers in the human gut, e.g., members of the *Christensenellales* order, as well as a core group of bacteria favoring its colonization of the gut environment. Syntrophy may underlie the identified associations, with potential effects on bacterial fermentation and so, on the human host. In addition, endoR, my method for interpreting machine learning models, applies to all sorts of problems being studied with tree ensemble models. Thus, the application of endoR is not limited to the microbiome field and will hopefully appear useful to other researchers investigating complex systems in the future. Furthermore, it could help deciphering host-microbe interactions occurring across all human microbiomes. Ultimately, comprehending dynamics of human microbial inhabitants will allow us to understand how these microorganisms, accounting for half of our cells [204], affect us and are part of us.

Bibliography

[1] S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methé, P. D. Schloss, D. Gevers, M. Mitreva, and C. Huttenhower. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLoS Computational Biology*, 8(6):e1002358, 6 2012. doi: 10.1371/journal.pcbi.1002358.

[2] D. Ai, H. Pan, R. Han, X. Li, G. Liu, and L. C. Xia. Using decision tree aggregation with random forest model to identify gut microbes associated with colorectal cancer. *Genes*, 10(2), 2019. doi: 10.3390/genes10020112.

[3] J. Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982. doi: 10.1111/j.2517-6161.1982.tb01195.x.

[4] G. Al-Ghalith and D. Knights. BURST enables optimal exhaustive DNA alignment for big data. 2017. doi: doi.org/10.5281/zenodo.806850.

[5] D. Albanese, C. De Filippo, D. Cavalieri, and C. Donati. Explaining Diversity in Metagenomic Datasets by Phylogenetic-Based Feature Weighting. *PLOS Computational Biology*, 11(3):e1004186, 3 2015. doi: 10.1371/journal.pcbi.1004186.

[6] A. Almeida, A. L. Mitchell, M. Boland, S. C. Forster, G. B. Gloor, A. Tarkowska, T. D. Lawley, and R. D. Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, 4 2019. doi: 10.1038/s41586-019-0965-1.

[7] A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz, N. Segata, N. C. Kyrpides, and R. D. Finn. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, 39(1):105–114, 1 2021. doi: 10.1038/s41587-020-0603-3.

[8] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 5 2010. doi: 10.1093/bioinformatics/btq134.

[9] F. Armougom, M. Henry, B. Vialettes, D. Raccah, and D. Raoult. Monitoring bacterial community of human gut microbiota reveals an increase in Lactobacillus in obese patients and Methanogens in anorexic patients. *PloS one*, 4(9):e7125, 9 2009. doi: 10.1371/journal.pone.0007125.

[10] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. G. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. M. de Vos, S. Brunak, J. Doré, J. Weissenbach, S. D. Ehrlich, and P. Bork. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 5 2011. doi: 10.1038/nature09944.

[11] L. Bajer, M. Kverka, M. Kostovcik, P. Macinga, J. Dvorak, Z. Stehlikova, J. Brezina, P. Wohl, J. Spicak, and P. Drastich. Distinct gut microbiota profiles in patients with primary sclerosing cholangitis and ulcerative colitis. *World Journal of Gastroenterology*, 23(25):4548, 2017. doi: 10.3748/wjg.v23.i25.4548.

[12] W. E. Balch and R. S. Wolfe. New Approach to the Cultivation of Methanogenic Bacteria: 2-Mercaptoethanesulfonic Acid (HS-CoM)-Dependent Growth of Methanobacterium ruminantium in

a Pressurized Atmosphere. *APPLiED AND ENVIRONMENTAL MICROBIOLOGY*, 32(6):781–791, 1976.

[13] C. Bang and R. A. Schmitz. Archaea associated with human surfaces: not to be underestimated. *FEMS Microbiology Reviews*, 39(5):631–648, 9 2015. doi: 10.1093/femsre/fuv010.

[14] S. Basu, K. Kumbier, J. B. Brown, and B. Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2 2018. doi: 10.1073/PNAS.1711236115.

[15] F. Beghini, E. Pasolli, T. D. Truong, L. Putignani, S. M. Cacciò, and N. Segata. Large-scale comparative metagenomics of Blastocystis, a common member of the human gut microbiome. *The ISME Journal*, 11(12):2848–2863, 12 2017. doi: 10.1038/ismej.2017.139.

[16] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.

[17] P. Biecek. DALEX: Explainers for Complex Predictive Models in R. *The Journal of Machine Learning Research*, 19(1):3245–3249, 2018.

[18] P. Biecek and T. Burzykowski. Explanatory Model Analysis, 2020. URL `https://pbiecek.github.io/ema/preface.html`.

[19] M. J. Bonder, A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann, A. V. Vila, P. Deelen, T. Vatanen, M. Schirmer, S. P. Smeekens, D. V. Zhernakova, S. A. Jankipersadsing, M. Jaeger, M. Oosting, M. C. Cenit, A. A. Masclee, M. A. Swertz, Y. Li, V. Kumar, L. Joosten, H. Harmsen, R. K. Weersma, L. Franke, M. H. Hofker, R. J. Xavier, D. Jonkers, M. G. Netea, C. Wijmenga, J. Fu, and A. Zhernakova. The effect of host genetics on the gut microbiome. *Nature Genetics*, 48(11):1407–1412, 2016. doi: 10.1038/ng.3663.

[20] D. R. Boone, R. L. Johnson, and Y. Liu. Diffusion of the Interspecies Electron Carriers H(2) and Formate in Methanogenic

Ecosystems and Its Implications in the Measurement of K(m) for H(2) or Formate Uptake. *Applied and environmental microbiology*, 55(7):1735–41, 7 1989.

[21] G. Borrel, H. M. B. Harris, W. Tottey, A. Mihajlovski, N. Parisot, E. Peyretaillade, P. Peyret, S. Gribaldo, P. W. O'Toole, and J.-F. Brugere. Genome Sequence of "Candidatus Methanomethylophilus alvus" Mx1201, a Methanogenic Archaeon from the Human Gut Belonging to a Seventh Order of Methanogens. *Journal of Bacteriology*, 194(24):6944–6945, 12 2012. doi: 10.1128/JB.01867-12.

[22] G. Borrel, H. M. B. Harris, N. Parisot, N. Gaci, W. Tottey, A. Mihajlovski, J. Deane, S. Gribaldo, O. Bardot, E. Peyretaillade, P. Peyret, P. W. O'Toole, and J.-F. Brugere. Genome Sequence of "Candidatus Methanomassiliicoccus intestinalis" Issoire-Mx1, a Third Thermoplasmatales-Related Methanogenic Archaeon from Human Feces. *Genome Announcements*, 1(4), 7 2013. doi: 10.1128/genomeA.00453-13.

[23] G. Borrel, A. McCann, J. Deane, M. C. Neto, D. B. Lynch, J.-F. Brugère, and P. W. O'Toole. Genomics and metagenomics of trimethylamine-utilizing Archaea in the human gut microbiome. *The ISME Journal*, 11(9):2059–2074, 9 2017. doi: 10.1038/ismej.2017.72.

[24] G. Borrel, J.-F. Brugère, S. Gribaldo, R. A. Schmitz, and C. Moissl-Eichinger. The host-associated archaeome. *Nature Reviews Microbiology*, 18(11):622–636, 11 2020. doi: 10.1038/s41579-020-0407-y.

[25] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. doi: 10.1201/9780429469275-8.

[26] L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001. doi: 10.1214/ss/1009213726.

[27] L. Breiman and A. Cutler. Manual on Setting Up, Using, and Understanding Random Forests, ver. 4.0, 2003. URL `https://www.stat.berkeley.edu/%5C~breiman/RandomForests/cc_home.htm`.

[28] A. Camara, S. Konate, M. Tidjani Alou, A. Kodio, A. H. Togo, S. Cortaredona, B. Henrissat, M. A. Thera, O. K. Doumbo, D. Raoult, and M. Million. Clinical evidence of the role of Methanobrevibacter smithii in severe acute malnutrition. *Scientific Reports*, 11(1):5426, 12 2021. doi: 10.1038/s41598-021-84641-8.

[29] P. D. Cani, M. Van Hul, C. Lefort, C. Depommier, M. Rastelli, and A. Everard. Microbial regulation of organismal energy homeostasis. *Nature Metabolism*, 1(1):34–46, 2019. doi: 10.1038/s42255-018-0017-4.

[30] C. Chassard and A. Bernalier-Donadille. H2 and acetate transfers during xylan fermentation between a butyrate-producing xylanolytic species and hydrogenotrophic microorganisms from the human gut. *FEMS Microbiology Letters*, 254(1):116–122, 1 2006. doi: 10.1111/j.1574-6968.2005.00016.x.

[31] C. Chassard, E. Delmas, C. Robert, and A. Bernalier-Donadille. The cellulose-degrading microbial community of the human gut varies according to the presence or absence of methanogens. *FEMS Microbiology Ecology*, 74(1):205–213, 10 2010. doi: 10.1111/j.1574-6941.2010.00941.x.

[32] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, 11 2019. doi: 10.1093/bioinformatics/btz848.

[33] C. Chen, A. Liaw, and L. Breiman. Using Random Forest to Learn Imbalanced Data. *University of California, Berkeley*, 110 (1-12):24, 2004.

[34] E. Z. Chen and H. Li. A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics*, 32(17):2611–2617, 9 2016. doi: 10.1093/bioinformatics/btw308.

[35] T. Chen and C. Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA,

8 2016. ACM. ISBN 9781450342322. doi: 10.1145/2939672.2939785.

[36] Y.-r. Chen, H.-m. Zheng, G.-x. Zhang, F.-l. Chen, L.-d. Chen, and Z.-c. Yang. High Oscillospira abundance indicates constipation and low BMI in the Guangdong Gut Microbiome Project. *Scientific Reports*, 10(1):9364, 12 2020. doi: 10.1038/s41598-020-66369-z.

[37] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[38] P. I. Costea, F. Hildebrand, M. Arumugam, F. Bäckhed, M. J. Blaser, F. D. Bushman, W. M. de Vos, S. D. Ehrlich, C. M. Fraser, M. Hattori, C. Huttenhower, I. B. Jeffery, D. Knights, J. D. Lewis, R. E. Ley, H. Ochman, P. W. O'Toole, C. Quince, D. A. Relman, F. Shanahan, S. Sunagawa, J. Wang, G. M. Weinstock, G. D. Wu, G. Zeller, L. Zhao, J. Raes, R. Knight, and P. Bork. Enterotypes in the landscape of gut microbial community composition. *Nature Microbiology*, 3(1):8–16, 1 2018. doi: 10.1038/s41564-017-0072-8.

[39] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Sy:1695, 2006.

[40] L. A. David, A. Weil, E. T. Ryan, S. B. Calderwood, J. B. Harris, F. Chowdhury, Y. Begum, F. Qadri, R. C. LaRocque, and P. J. Turnbaugh. Gut microbial succession follows acute secretory diarrhea in humans. *mBio*, 6(3):00381–15, 5 2015. doi: 10.1128/mBio.00381-15.

[41] J. De La Cuesta-Zuluaga, N. T. Mueller, V. Corrales-Agudelo, E. P. Vel??squez-Mej??a, J. A. Carmona, J. M. Abad, and J. S. Escobar. Metformin is associated with higher relative abundance of mucin-degrading akkermansia muciniphila and several short-chain fatty acid-producing microbiota in the gut. *Diabetes Care*, 40(1):54–62, 2017. doi: 10.2337/dc16-1324.

[42] J. de la Cuesta-Zuluaga, V. Corrales-Agudelo, E. P. Velásquez-Mejía, J. A. Carmona, J. M. Abad, and J. S. Escobar. Gut microbiota is associated with obesity and cardiometabolic disease

in a population in the midst of Westernization. *Scientific Reports*, 8(1):11356, 12 2018. doi: 10.1038/s41598-018-29687-x.

[43] J. de la Cuesta-Zuluaga, R. E. Ley, and N. D. Youngblut. Struo: a pipeline for building custom databases for common metagenome profilers. *Bioinformatics*, 36(7):2314–2315, 4 2020. doi: 10.1093/bioinformatics/btz899.

[44] J. de la Cuesta-Zuluaga, T. D. Spector, N. D. Youngblut, and R. E. Ley. Genomic Insights into Adaptations of Trimethylamine-Utilizing Methanogens to Diverse Habitats, Including the Human Gut. *mSystems*, 6(1), 2 2021. doi: 10.1128/mSystems.00939-20.

[45] F. Degenhardt, S. Seifert, and S. Szymczak. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, 20(2):492–503, 2019. doi: 10.1093/bib/bbx124.

[46] G. den Besten, K. van Eunen, A. K. Groen, K. Venema, D.-J. Reijngoud, and B. M. Bakker. The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *Journal of lipid research*, 54(9):2325–40, 9 2013. doi: 10.1194/jlr.R036012.

[47] H. Deng. Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, 7(4):277–287, 2019. doi: 10.1007/s41060-018-0144-8.

[48] H. Deng and G. Runger. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12):3483–3489, 2013. doi: 10.1016/j.patcog.2013.05.018.

[49] F. E. Dewhirst, T. Chen, J. Izard, B. J. Paster, A. C. R. Tanner, W.-H. Yu, A. Lakshmanan, and W. G. Wade. The human oral microbiome. *Journal of bacteriology*, 192(19):5002–5017, 2010.

[50] P. Domingos. MetaCost. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, pages 155–164, New York, New York, USA, 1999. ACM Press. ISBN 1581131437. doi: 10.1145/312129.312220.

[51] P. Domingos. A Few Useful Things to Know About Machine Learning. *communications of the ACM*, 55(10):79–88, 2012.

[52] M. Dowle and A. Srinivasan. data.table: Extension of ‘data.frame‘, 2020.

[53] N. R. Draper and H. Smith. *Applied regression analysis*. 1998. ISBN 0-471-17082-8.

[54] B. Dridi, M. Henry, A. El Khéchine, D. Raoult, and M. Drancourt. High Prevalence of Methanobrevibacter smithii and Methanosphaera stadtmanae Detected in the Human Gut Using an Improved DNA Detection Protocol. *PLoS ONE*, 4(9):e7063, 9 2009. doi: 10.1371/journal.pone.0007063.

[55] B. Dridi, D. Raoult, and M. Drancourt. Archaea as emerging organisms in complex human microbiomes. *Anaerobe*, 17(2):56–63, 4 2011. doi: 10.1016/J.ANAEROBE.2011.03.001.

[56] B. Dridi, M.-L. Fardeau, B. Ollivier, D. Raoult, and M. Drancourt. Methanomassiliicoccus luminyensis gen. nov., sp. nov., a methanogenic archaeon isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology*, 62(Pt_8): 1902–1907, 8 2012. doi: 10.1099/ijs.0.033712-0.

[57] C. Duvallet, S. M. Gibbons, T. Gurry, R. A. Irizarry, and E. J. Alm. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nature Communications*, 8(1), 2017. doi: 10.1038/s41467-017-01973-8.

[58] D. Eddelbuettel and R. François. Rcpp : Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8), 2011. doi: 10.18637/jss.v040.i08.

[59] Q. Feng, S. Liang, H. Jia, A. Stadlmayr, L. Tang, Z. Lan, D. Zhang, H. Xia, X. Xu, Z. Jie, L. Su, X. Li, X. Li, J. Li, L. Xiao, U. Huber-Schönauer, D. Niederseer, X. Xu, J. Y. Al-Aama, H. Yang, J. Wang, K. Kristiansen, M. Arumugam, H. Tilg, C. Datz, and J. Wang. Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature Communications*, 6(1):6528, 12 2015. doi: 10.1038/ncomms7528.

[60] D. Feria-Gervasio, W. Tottey, N. Gaci, M. Alric, J.-M. Cardot, P. Peyret, J.-F. Martin, E. Pujos, J.-L. Sébédio, and J.-F. Brugère. Three-stage continuous culture system with a self-generated anaerobia to study the regionalized metabolism of the human gut microbiota. *Journal of Microbiological Methods*, 96: 111–118, 2014. doi: 10.1016/j.mimet.2013.11.015.

[61] J. Fernandes, A. Wang, W. Su, S. R. Rozenbloom, A. Taibi, E. M. Comelli, and T. M. S. Wolever. Age, dietary fiber, breath methane, and fecal short chain fatty acids are interrelated in Archaea-positive humans. *The Journal of nutrition*, 143(8):1269–75, 8 2013. doi: 10.3945/jn.112.170894.

[62] J. G. Ferry, editor. *Methanogenesis.* Springer US, Boston, MA, 1993. ISBN 978-1-4613-6013-1. doi: 10.1007/978-1-4615-2391-8.

[63] H. J. Flint, K. P. Scott, S. H. Duncan, P. Louis, and E. Forano. Microbial degradation of complex carbohydrates in the gut. *Gut microbes*, 3:4(August):289–306, 2012.

[64] H. J. Flint, S. H. Duncan, K. P. Scott, and P. Louis. Links between diet, gut microbiota composition and gut metabolism. *Proceedings of the Nutrition Society*, 74(01):13–22, 2 2015. doi: 10.1017/S0029665114001463.

[65] K. Forslund, F. Hildebrand, T. Nielsen, G. Falony, E. L. Chatelier, S. Sunagawa, E. Prifti, S. Vieira-Silva, V. Gudmundsdottir, H. Krogh Pedersen, M. Arumugam, K. Kristiansen, A. Y. Voigt, H. Vestergaard, R. Hercog, P. I. Costea, J. R. Kultima, J. Li, T. Jørgensen, F. Levenez, J. Dore, M. Consortium, H. B. Nielsen, S. Brunak, J. Raes, T. Hansen, and J. Wang. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. *Nature*, 528:262–266, 2015. doi: 10.1038/nature15766.

[66] S. C. Forster, N. Kumar, B. O. Anonye, A. Almeida, E. Viciani, M. D. Stares, M. Dunn, T. T. Mkandawire, A. Zhu, Y. Shao, L. J. Pike, T. Louie, H. P. Browne, A. L. Mitchell, B. A. Neville, R. D. Finn, and T. D. Lawley. A human gut

bacterial genome and culture collection for improved metagenomic analyses. *Nature Biotechnology*, 37(2):186–192, 2 2019. doi: 10.1038/s41587-018-0009-7.

[67] E. A. Franzosa, L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata, and C. Huttenhower. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15(11):962–968, 11 2018. doi: 10.1038/s41592-018-0176-y.

[68] J. Friedman and B. Popescu. Importance sampled learning ensembles. *Journal of Machine Learning Research*, 94305(2):1–32, 2003.

[69] J. Friedman, T. Hastie, and R. Tibshirani. Sparse covariance estimation. *Biostatistics*, 9(3):432–441, 2008.

[70] J. Friedman, E. J. Alm, S. Westcott, E. Cosgrove, and B. Hayete. Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*, 8(9):e1002687, 9 2012. doi: 10.1371/journal.pcbi.1002687.

[71] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 10 2001. doi: 10.1214/aos/1013203451.

[72] J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3):916–954, 2008. doi: 10.1214/07-AOAS148.

[73] C. Frioux, D. Singh, T. Korcsmaros, and F. Hildebrand. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. *Computational and Structural Biotechnology Journal*, 18:1722–1734, 2020. doi: 10.1016/j.csbj.2020.06.028.

[74] J. Fu, M. J. Bonder, M. C. Cenit, E. F. Tigchelaar, A. Maatman, J. A. Dekens, E. Brandsma, J. Marczynska, F. Imhann, R. K. Weersma, L. Franke, T. W. Poon, R. J. Xavier, D. Gevers, M. H. Hofker, C. Wijmenga, and A. Zhernakova. The Gut Microbiome Contributes to a Substantial Proportion of the Variation in Blood

Lipids. *Circulation Research*, 117(9):817–824, 10 2015. doi: 10.1161/CIRCRESAHA.115.306807.

[75] J. C. Garcia-Betancur, A. Yepes, J. Schneider, and D. Lopez. Single-cell analysis of Bacillus subtilis biofilms using fluorescence microscopy and flow cytometry. *Journal of visualized experiments : JoVE*, (60):1–8, 2 2012. doi: 10.3791/3796.

[76] J. L. Gehrig, S. Venkatesh, H.-W. Chang, M. C. Hibberd, V. L. Kung, J. Cheng, R. Y. Chen, S. Subramanian, C. A. Cowardin, M. F. Meier, D. O'Donnell, M. Talcott, L. D. Spears, C. F. Semenkovich, B. Henrissat, R. J. Giannone, R. L. Hettich, O. Ilkayeva, M. Muehlbauer, C. B. Newgard, C. Sawyer, R. D. Head, D. A. Rodionov, A. A. Arzamasov, S. A. Leyn, A. L. Osterman, M. I. Hossain, M. Islam, N. Choudhury, S. A. Sarker, S. Huq, I. Mahmud, I. Mostafa, M. Mahfuz, M. J. Barratt, T. Ahmed, and J. I. Gordon. Effects of microbiota-directed foods in gnotobiotic animals and undernourished children. *Science*, 365(6449): eaau4732, 7 2019. doi: 10.1126/science.aau4732.

[77] S. R. Gill, M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson. Metagenomic Analysis of the Human Distal Gut Microbiome. *Science*, 312(5778):1355–1359, 6 2006. doi: 10.1126/science.1124234.

[78] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8(November):1–6, 11 2017. doi: 10.3389/fmicb.2017.02224.

[79] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 1 2015. doi: 10.1080/10618600.2014.907095.

[80] J. K. Goodrich, J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren, R. Blekhman, M. Beaumont, W. Van Treuren, R. Knight, J. T. Bell, T. D. Spector, A. G. Clark, R. E. Ley, W. V. Treuren,

R. Knight, J. T. Bell, T. D. Spector, A. G. Clark, and R. E. Ley. Human genetics shape the gut microbiome. *Cell*, 6(1594): 789–79909, 2014. doi: 10.1016/j.cell.2014.09.053.

[81] J. K. Goodrich, E. R. Davenport, M. Beaumont, M. A. Jackson, R. Knight, C. Ober, T. D. Spector, J. T. Bell, A. G. Clark, and R. E. Ley. Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host & Microbe*, 19(5):731–743, 5 2016. doi: 10.1016/j.chom.2016.04.017.

[82] R. D. Goodrich, J. E. Garrett, D. R. Gast, M. A. Kirick, D. A. Larson, and J. C. Meiske. Influence of Monensin on the Performance of Cattle. *Journal of Animal Science*, 58(6):1484–1498, 6 1984. doi: 10.2527/jas1984.5861484x.

[83] U. Gophna, T. Konikoff, and H. B. Nielsen. Oscillospira and related bacteria – From metagenomic species to metabolic features. *Environmental Microbiology*, 19(3):835–841, 2017. doi: 10.1111/1462-2920.13658.

[84] A. Gosiewska and P. Biecek. Do Not Trust Additive Explanations. *arXiv*, 3 2019.

[85] W. Gou, C.-w. Ling, Y. He, Z. Jiang, Y. Fu, F. Xu, Z. Miao, T.-y. Sun, J.-s. Lin, H.-l. Zhu, H. Zhou, Y.-m. Chen, and J.-S. Zheng. Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes. *Diabetes Care*, (1):dc201536, 2020. doi: 10.2337/dc20-1536.

[86] W. Gou, C.-w. Ling, Y. He, Z. Jiang, Y. Fu, F. Xu, Z. Miao, T.-y. Sun, J.-s. Lin, H.-l. Zhu, H. Zhou, Y.-m. Chen, and J.-S. Zheng. Interpretable Machine Learning Framework Reveals Robust Gut Microbiome Features Associated With Type 2 Diabetes. *Diabetes Care*, 44(2):358–366, 2 2021. doi: 10.2337/dc20-1536.

[87] N. S. Grantham, Y. Guan, B. J. Reich, E. T. Borer, and K. Gross. MIMIX: A Bayesian Mixed-Effects Model for Microbiome Data From Designed Experiments. *Journal of the American Statistical Association*, 115(530):599–609, 4 2020. doi: 10.1080/01621459.2019.1626242.

[88] B. Greenwell, B. Boehmke, J. Cunningham, and G. Developers. gbm: Generalized Boosted Regression Models, 2020. URL `https://cran.r-project.org/package=gbm`.

[89] L. L. Grønkjær. Periodontal disease and liver cirrhosis: A systematic review. *SAGE open medicine*, 3:2050312115601122, 2015.

[90] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine learning*, 46(1):389–422, 2002. doi: 10.1023{_}A1012487302797.

[91] J. Halfvarson, C. J. Brislawn, R. Lamendella, Y. Vázquez-Baeza, W. A. Walters, L. M. Bramer, M. D'Amato, F. Bonfiglio, D. McDonald, A. Gonzalez, E. E. McClure, M. F. Dunklebarger, R. Knight, and J. K. Jansson. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature Microbiology*, 2 (5):17004, 5 2017. doi: 10.1038/nmicrobiol.2017.4.

[92] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–R249, 10 1998. doi: 10.1016/S1074-5521(98)90108-9.

[93] E. E. Hansen, C. A. Lozupone, F. E. Rey, M. Wu, J. L. Guruge, A. Narra, J. Goodfellow, J. R. Zaneveld, D. T. McDonald, J. A. Goodrich, A. C. Heath, R. Knight, and J. I. Gordon. Supporting Information - Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proceedings of the National Academy of Sciences*, 2011. doi: 10.1073/pnas.1000071108.

[94] E. E. Hansen, C. A. Lozupone, F. E. Rey, M. Wu, J. L. Guruge, A. Narra, J. Goodfellow, J. R. Zaneveld, D. T. Mcdonald, J. A. Goodrich, A. C. Heath, R. Knight, J. I. Gordon, and T. R. Klaenhammer. Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proceedings of the National Academy of Sciences*, 2011. doi: 10.1073/pnas.1000071108.

[95] F. E. Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

[96] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009. doi: 10.1111/j.1751-5823.2009.00095{_}18.x.

[97] C. Hoffmann, S. Dollive, S. Grunberg, J. Chen, H. Li, G. D. Wu, J. D. Lewis, and F. D. Bushman. Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PLoS ONE*, 8(6):e66019, 6 2013. doi: 10.1371/journal.pone.0066019.

[98] I. Holmes, K. Harris, and C. Quince. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE*, 7(2), 2012. doi: 10.1371/journal.pone.0030126.

[99] T. Hooven, Y. C. Lin, and A. Salleb-Aouissi. Multiple instance learning for predicting necrotizing enterocolitis in premature infants using microbiome data. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 99–109, New York, NY, USA, 4 2020. ACM. ISBN 9781450370462. doi: 10.1145/3368555.3384466.

[100] Houtao Deng and G. Runger. Feature selection via regularized trees. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 6 2012. ISBN 978-1-4673-1490-9. doi: 10.1109/IJCNN.2012.6252640.

[101] T. Hu, P. Gallins, and Y.-H. Zhou. A zero-inflated beta-binomial model for microbiome data analysis. *Stat*, 7(1):e185, 2018. doi: 10.1002/sta4.185.

[102] F. Hugenholtz, J. A. Mullaney, M. Kleerebezem, H. Smidt, and D. I. Rosendale. Modulation of the microbial fermentation in the gut by fermentable carbohydrates. *Bioactive Carbohydrates and Dietary Fibre*, 2(2):133–142, 2013. doi: 10.1016/j.bcdf.2013.09.008.

[103] P. Hugenholtz. Exploring prokaryotic diversity in the genomic era. *Genome Biology*, 3(2):1–8, 2002. doi: 10.1186/gb-2002-3-2-reviews0003.

[104] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386, 2 2007. doi: 10.1101/gr.5969107.

[105] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–1560, 9 2011. doi: 10.1101/gr.120618.111.

[106] A. Ignacio, M. Fernandes, V. Rodrigues, F. Groppo, A. Cardoso, M. Avila-Campos, and V. Nakano. Correlation between body mass index and faecal microbiota from children. *Clinical Microbiology and Infection*, 22(3):1–258, 3 2016. doi: 10.1016/j.cmi.2015.10.031.

[107] S. Janitza, E. Celik, and A.-L. Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, pages 1–31, 11 2016. doi: 10.1007/s11634-016-0276-4.

[108] H. Jiang, R. Lei, S.-W. Ding, and S. Zhu. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics*, 15(1):182, 12 2014. doi: 10.1186/1471-2105-15-182.

[109] Y. Jiang, P. Biecek, O. Paluszyńska, Agasitko, and K. Kobylinska. ModelOriented/randomForestExplainer: CRAN release 0.10.1, 2020. URL `https://doi.org/10.5281/zenodo.3941250`.

[110] Z. Jie, H. Xia, S.-L. Zhong, Q. Feng, S. Li, S. Liang, H. Zhong, Z. Liu, Y. Gao, H. Zhao, D. Zhang, Z. Su, Z. Fang, Z. Lan, J. Li, L. Xiao, J. Li, R. Li, X. Li, F. Li, H. Ren, Y. Huang, Y. Peng, G. Li, B. Wen, B. Dong, J.-Y. Chen, Q.-S. Geng, Z.-W. Zhang, H. Yang, J. Wang, J. Wang, X. Zhang, L. Madsen, S. Brix, G. Ning, X. Xu, X. Liu, Y. Hou, H. Jia, K. He,

and K. Kristiansen. The gut microbiome in atherosclerotic cardiovascular disease. *Nature Communications*, 8(1), 2017. doi: 10.1038/s41467-017-00900-1.

[111] K. A. Johnson and D. E. Johnson. Methane emissions from cattle. *Journal of Animal Science*, 73(8):2483–2492, 8 1995. doi: 10.2527/1995.7382483x.

[112] D. D. Kang, J. Froula, R. Egan, and Z. Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 8 2015. doi: 10.7717/peerj.1165.

[113] S. Khelaifia, J.-C. Lagier, V. D. Nkamga, E. Guilhot, M. Drancourt, and D. Raoult. Aerobic culture of methanogenic archaea without an external source of hydrogen. *European Journal of Clinical Microbiology & Infectious Diseases*, 35(6):985–991, 6 2016. doi: 10.1007/s10096-016-2627-7.

[114] N. Klimenko, A. Tyakht, A. Popenko, A. Vasiliev, I. Altukhov, D. Ischenko, T. Shashkova, D. Efimova, D. Nikogosov, D. Osipenko, S. Musienko, K. Selezneva, A. Baranova, A. Kurilshikov, S. Toshchakov, A. Korzhenkov, N. Samarov, M. Shevchenko, A. Tepliuk, and D. Alexeev. Microbiome Responses to an Uncontrolled Short-Term Diet Intervention in the Frame of the Citizen Science Project. *Nutrients*, 10(5):576, 5 2018. doi: 10.3390/nu10050576.

[115] R. Knight, A. Vrbanac, B. C. Taylor, A. Aksenov, C. Callewaert, J. Debelius, A. Gonzalez, T. Kosciolek, L. I. McCall, D. McDonald, A. V. Melnik, J. T. Morton, J. Navas, R. A. Quinn, J. G. Sanders, A. D. Swafford, L. R. Thompson, A. Tripathi, Z. Z. Xu, J. R. Zaneveld, Q. Zhu, J. G. Caporaso, and P. C. Dorrestein. Best practices for analysing microbiomes, 7 2018. ISSN 17401534. URL `http://www.nature.com/articles/s41579-018-0029-9`.

[116] D. Knights, E. K. Costello, and R. Knight. Supervised classification of human microbiota. *FEMS Microbiology Reviews*, 35(2): 343–359, 2011. doi: 10.1111/j.1574-6976.2010.00251.x.

[117] J. E. Koenig, A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent, R. E. Ley, and T. R. Klaenhammer. Succession of microbial consortia in the developing infant gut microbiome. *PNAS*, 108:4578–4585, 2011. doi: 10.1073/pnas.1000081107.

[118] R. Kohn and R. Boston. The role of thermodynamics in controlling rumen metabolism. In J. P. McNamara, J. France, and D. E. Beever, editors, *Modelling nutrient utilization in farm animals*, pages 11–24. CAB International, 2000. ISBN 0 85199 449 0.

[119] O. Koren, D. Knights, A. Gonzalez, L. Waldron, N. Segata, R. Knight, C. Huttenhower, and R. E. Ley. A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLoS Computational Biology*, 9(1):e1002863, 1 2013. doi: 10.1371/journal.pcbi.1002863.

[120] R. Krajmalnik-Brown, Z.-E. Ilhan, D.-W. Kang, and J. K. Dibaise. Effects of Gut Microbes on Nutrient Absorption and Energy Regulation. *Nutrition in Clinical Practice*, 27(2):201–214, 2012. doi: 10.1177/0884533611436116.

[121] D. Kunkel, R. J. Basseri, M. D. Makhani, K. Chong, C. Chang, and M. Pimentel. Methane on Breath Testing Is Associated with Constipation: A Systematic Review and Meta-analysis. *Digestive Diseases and Sciences*, 56(6):1612–1618, 6 2011. doi: 10.1007/s10620-011-1590-5.

[122] A. Kurilshikov, C. Medina-Gomez, R. Bacigalupe, D. Radjabzadeh, J. Wang, A. Demirkan, C. I. Le Roy, J. A. Raygoza Garay, C. T. Finnicum, X. Liu, D. V. Zhernakova, M. J. Bonder, T. H. Hansen, F. Frost, M. C. Rühlemann, W. Turpin, J.-Y. Moon, H.-N. Kim, K. Lüll, E. Barkan, S. A. Shah, M. Fornage, J. Szopinska-Tokov, Z. D. Wallen, D. Borisevich, L. Agreus, A. Andreasson, C. Bang, L. Bedrani, J. T. Bell, H. Bisgaard, M. Boehnke, D. I. Boomsma, R. D. Burk, A. Claringbould, K. Croitoru, G. E. Davies, C. M. van Duijn, L. Duijts, G. Falony, J. Fu, A. van der Graaf, T. Hansen, G. Homuth, D. A. Hughes, R. G. Ijzerman, M. A. Jackson, V. W. V. Jaddoe, M. Joossens,

T. Jørgensen, D. Keszthelyi, R. Knight, M. Laakso, M. Laudes, L. J. Launer, W. Lieb, A. J. Lusis, A. A. M. Masclee, H. A. Moll, Z. Mujagic, Q. Qibin, D. Rothschild, H. Shin, S. J. Sørensen, C. J. Steves, J. Thorsen, N. J. Timpson, R. Y. Tito, S. Vieira-Silva, U. Völker, H. Völzke, U. Võsa, K. H. Wade, S. Walter, K. Watanabe, S. Weiss, F. U. Weiss, O. Weissbrod, H.-J. Westra, G. Willemsen, H. Payami, D. M. A. E. Jonkers, A. Arias Vasquez, E. J. C. de Geus, K. A. Meyer, J. Stokholm, E. Segal, E. Org, C. Wijmenga, H.-L. Kim, R. C. Kaplan, T. D. Spector, A. G. Uitterlinden, F. Rivadeneira, A. Franke, M. M. Lerch, L. Franke, S. Sanna, M. D'Amato, O. Pedersen, A. D. Paterson, R. Kraaij, J. Raes, and A. Zhernakova. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nature Genetics*, 53(2):156–165, 2 2021. doi: 10.1038/s41588-020-00763-1.

[123] M. B. Kursa and W. R. Rudnicki. Feature Selection with the Boruta Package. *Journal Of Statistical Software*, 36(11):1–13, 2010. doi: Vol.36,Issue11,Sep2010.

[124] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Computational Biology*, 11(5):1–25, 2015. doi: 10.1371/journal.pcbi.1004226.

[125] J. Lambrecht, N. Cichocki, T. Hübschmann, C. Koch, H. Harms, and S. Müller. Flow cytometric quantification, sorting and sequencing of methanogenic archaea based on F420 autofluorescence. *Microbial cell factories*, 16(1):180, 10 2017. doi: 10.1186/s12934-017-0793-7.

[126] J. R. Landis and G. G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159, 3 1977. doi: 10.2307/2529310.

[127] S. K. P. Lau, A. McNabb, G. K. S. Woo, L. Hoang, A. M. Y. Fung, L. M. W. Chung, P. C. Y. Woo, and K.-Y. Yuen. Catabacter hongkongensis gen. nov., sp. nov., Isolated from Blood Cultures of Patients from Hong Kong and Canada. *Journal of Clinical Microbiology*, 45(2):395–401, 2 2007. doi: 10.1128/JCM.01831-06.

[128] S. C. Leahy, W. J. Kelly, E. Altermann, R. S. Ronimus, C. J. Yeoman, D. M. Pacheco, D. Li, Z. Kong, S. McTavish, C. Sang, S. C. Lambie, P. H. Janssen, D. Dey, and G. T. Attwood. The Genome Sequence of the Rumen Methanogen Methanobrevibacter ruminantium Reveals New Possibilities for Controlling Ruminant Methane Emissions. *PLoS ONE*, 5(1):e8926, 1 2010. doi: 10.1371/journal.pone.0008926.

[129] A. Liaw, M. Wiener, and Others. Classification and regression by randomForest. *R news*, 2(3):18–22, 2002.

[130] M. Y. Lim, H. J. You, H. S. Yoon, B. Kwon, J. Y. Lee, S. Lee, Y. M. Song, K. Lee, J. Sung, and G. Ko. The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. *Gut*, 66(6):1031–1038, 2017. doi: 10.1136/gutjnl-2015-311326.

[131] B. Liu, L. L. Faller, N. Klitgord, V. Mazumdar, M. Ghodsi, D. D. Sommer, T. R. Gibbons, T. J. Treangen, Y.-C. Chang, S. Li, and others. Deep sequencing of the oral microbiome reveals signatures of periodontal disease. *PloS one*, 7(6):e37919, 2012.

[132] X. Liu, J. L. Sutter, J. de la Cuesta-Zuluaga, J. L. Waters, N. D. Youngblut, and R. E. Ley. Reclassification of Catabacter hongkongensis as Christensenella hongkongensis comb. nov. based on whole genome analysis. *International Journal of Systematic and Evolutionary Microbiology*, 2021. doi: 10.1099/ijsem.0.004774.

[133] Y. Liu and A. Just. SHAPforxgboost: SHAP Plots for 'XGBoost', 2020. URL `https://github.com/liuyanguu/SHAPforxgboost/`.

[134] Y. Liu and W. B. Whitman. Metabolic, Phylogenetic, and Ecological Diversity of the Methanogenic Archaea. *Annals of the New York Academy of Sciences*, 1125(1):171–189, 3 2008. doi: 10.1196/annals.1419.019.

[135] M. Loftus, S. A. D. Hassouneh, and S. Yooseph. Bacterial associations in the healthy human gut microbiome across populations. *Scientific Reports*, 11(1):1–14, 2021. doi: 10.1038/s41598-021-82449-0.

[136] V. López, A. Fernández, S. García, V. Palade, and F. Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250:113–141, 11 2013. doi: 10.1016/j.ins.2013.07.007.

[137] P. Louis and H. J. Flint. Formation of propionate and butyrate by the human colonic microbiota. *Environmental Microbiology*, 19(1):29–41, 1 2017. doi: 10.1111/1462-2920.13589.

[138] P. Louis, G. L. Hold, and H. J. Flint. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology*, 12(10):661–672, 10 2014. doi: 10.1038/nrmicro3344.

[139] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3:e104, 1 2017. doi: 10.7717/peerj-cs.104.

[140] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30*, pages 4765–4774, 2017.

[141] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K. W. Low, S. F. Newman, J. Kim, and S. I. Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749–760, 2018. doi: 10.1038/s41551-018-0304-0.

[142] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv*, 2019.

[143] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

[144] I. Mack, U. Cuntz, C. Grämer, S. Niedermaier, C. Pohl, A. Schwiertz, K. Zimmermann, S. Zipfel, P. Enck, and J. Penders. Weight gain in anorexia nervosa does not ameliorate the

faecal microbiota, branched chain fatty acid profiles, and gastrointestinal complaints. *Scientific reports*, 6:26752, 2016. doi: 10.1038/srep26752.

[145] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. cluster: Cluster Analysis Basics and Extensions, 2019.

[146] L. Maier, M. Pruteanu, M. Kuhn, G. Zeller, A. Telzerow, E. E. Anderson, A. R. Brochado, K. C. Fernandez, H. Dose, H. Mori, K. R. Patil, P. Bork, and A. Typas. Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature*, 555(7698):623–628, 2018. doi: 10.1038/nature25979.

[147] L. Mancabelli, C. Milani, G. A. Lugli, F. Turroni, C. Ferrario, D. van Sinderen, and M. Ventura. Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. *Environmental Microbiology*, 19(4):1379–1390, 4 2017. doi: 10.1111/1462-2920.13692.

[148] R. Mathur, K. S. Chua, M. Mamelak, W. Morales, G. M. Barlow, R. Thomas, D. Stefanovski, S. Weitsman, Z. Marsh, R. N. Bergman, and M. Pimentel. Metabolic effects of eradicating breath methane using antibiotics in prediabetic subjects with obesity. *Obesity*, 24(3):576–582, 3 2016. doi: 10.1002/oby.21385.

[149] W. Mazier, K. Le Corf, C. Martinez, H. Tudela, D. Kissi, C. Kropp, C. Coubard, M. Soto, F. Elustondo, G. Rawadi, and S. P. Claus. A New Strain of Christensenella minuta as a Potential Biotherapy for Obesity and Associated Metabolic Diseases. *Cells*, 10(4):823, 4 2021. doi: 10.3390/cells10040823.

[150] C. A. Mbakwa, J. Penders, P. H. Savelkoul, C. Thijs, P. C. Dagnelie, M. Mommers, and I. C. Arts. Gut colonization with methanobrevibacter smithii is associated with childhood weight development. *Obesity*, 23(12):2508–2516, 12 2015. doi: 10.1002/oby.21266.

[151] C. E. McCulloch and J. M. Neuhaus. Generalized Linear Mixed Models. *Encyclopedia of Biostatistics*, 2005. doi: 10.1002/0470011815.b2a10021.

[152] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3): 1436–1462, 2006. doi: 10.1214/009053606000000281.

[153] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 72(4):417–473, 2010. doi: 10.1111/j.1467-9868.2010.00740.x.

[154] A. Mihajlovski, J. Doré, F. Levenez, M. Alric, and J.-F. Brugère. Molecular evaluation of the human gut methanogenic archaeal microbiota reveals an age-associated increase of the diversity. *Environmental Microbiology Reports*, 2(2):272–280, 4 2010. doi: 10.1111/j.1758-2229.2009.00116.x.

[155] T. L. Miller. Methanobrevibacter. In W. B. Whitman, F. Rainey, P. Kämpfer, M. Trujillo, J. Chun, P. DeVos, B. Hedlund, and S. Dedysh, editors, *Bergey's Manual of Systematics of Archaea and Bacteria*, pages 1–14. John Wiley & Sons, Ltd, Chichester, UK, 9 2015. ISBN 9781118960608. doi: 10.1002/9781118960608.gbm00496.

[156] T. L. Miller and M. Wolin. Methanogens in human and animal intestinal Tracts. *Systematic and Applied Microbiology*, 7(2-3): 223–229, 5 1986. doi: 10.1016/S0723-2020(86)80010-8.

[157] M. Million, M. Maraninchi, M. Henry, F. Armougom, H. Richet, P. Carrieri, R. Valero, D. Raccah, B. Vialettes, and D. Raoult. Obesity-associated gut microbiota is enriched in Lactobacillus reuteri and depleted in Bifidobacterium animalis and Methanobrevibacter smithii. *International Journal of Obesity*, 36(6):817–825, 6 2012. doi: 10.1038/ijo.2011.153.

[158] C. Molnar, G. Casalicchio, and B. Bischl. iml: An R package for interpretable machine learning. *Journal of Open Source Software*, 3(26):786, 2018.

[159] M. Morotomi, F. Nagai, and Y. Watanabe. Description of Christensenella minuta gen. nov., sp. nov., isolated from human faeces, which forms a distinct branch in the order Clostridiales, and proposal of Christensenellaceae fam. nov. *INTERNA-*

TIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY, 62(1):144–149, 1 2012. doi: 10.1099/ijs.0.026989-0.

[160] B. E. L. Morris, R. Henneberger, H. Huber, and C. Moissl-eichinger. Microbial syntrophy : interaction for the common good. *FEMS Microbiology Reviews*, 37(3):384–406, 2013. doi: 10.1111/1574-6976.12019.

[161] D. J. Morrison and T. Preston. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut microbes*, 7(3):189–200, 5 2016. doi: 10.1080/19490976.2015.1134082.

[162] S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, and N. C. Kyrpides. New insights from uncultivated genomes of the global human gut microbiome. *Nature*, 568(7753):505–510, 4 2019. doi: 10.1038/s41586-019-1058-x.

[163] S. Ndongo, G. Dubourg, S. Khelaifia, P.-E. Fournier, and D. Raoult. Christensenella timonensis, a new bacterial species isolated from the human gut. *New Microbes and New Infections*, 13:32–33, 9 2016. doi: 10.1016/j.nmni.2016.05.010.

[164] S. Ndongo, S. Khelaifia, P.-E. Fournier, and D. Raoult. Christensenella massiliensis, a new bacterial species isolated from the human gut. *New Microbes and New Infections*, 12:69–70, 7 2016. doi: 10.1016/j.nmni.2016.04.014.

[165] F. Ng, S. Kittelmann, M. L. Patchett, G. T. Attwood, P. H. Janssen, J. Rakonjac, and D. Gagic. An adhesin from hydrogen-utilizing rumen methanogen M ethanobrevibacter ruminantium M1 binds a broad range of hydrogen-producing microorganisms. *Environmental Microbiology*, 18(9):3010–3021, 9 2016. doi: 10.1111/1462-2920.13155.

[166] J. K. Nicholson, E. Holmes, J. Kinross, R. Burcelin, G. Gibson, W. Jia, S. Pettersson, E. K. Costello, K. Stagaman, L. Dethlefsen, B. J. M. Bohannan, and D. A. Relman. Host-Gut Microbiota Metabolic Interactions. *Science (New York, N.Y.)*, 336(6086):1262–7, 6 2012. doi: 10.1126/science.1223813.

[167] H. B. Nielsen, M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. Le Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J.-M. Batto, M. B. Quintanilha dos Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Doré, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas, S. Kennedy, K. Kristiansen, J. R. Kultima, P. Léonard, F. Levenez, O. Lund, B. Moumen, D. Le Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sørensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, and S. D. Ehrlich. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, 32(8):822–828, 8 2014. doi: 10.1038/nbt.2939.

[168] S. Nishijima, W. Suda, K. Oshima, S.-W. Kim, Y. Hirose, H. Morita, and M. Hattori. Stability of human methanogenic flora over 35 years and a review of insights obtained from breath methane measurements. *Clin. Gastroenterol. Hepatol.*, 4(2):123–9, 4 2016. doi: 10.1093/dnares/dsw002.

[169] V. Nkamga, R. Lotte, P.-M. Roger, M. Drancourt, and R. Ruimy. Methanobrevibacter smithii and Bacteroides thetaiotaomicron cultivated from a chronic paravertebral muscle abscess. *Clinical Microbiology and Infection*, 22(12):1008–1009, 12 2016. doi: 10.1016/j.cmi.2016.09.007.

[170] A. J. Obregon-Tito, R. Y. Tito, J. Metcalf, K. Sankaranarayanan, J. C. Clemente, L. K. Ursell, Z. Zech Xu, W. Van Treuren, R. Knight, P. M. Gaffney, P. Spicer, P. Lawson, L. Marin-Reyes, O. Trujillo-Villarroel, M. Foster, E. Guija-Poma, L. Troncoso-Corzo, C. Warinner, A. T. Ozga, and C. M. Lewis. Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications*, 6(1):6505, 5 2015. doi: 10.1038/ncomms7505.

[171] T. G. Oh, S. M. Kim, C. Caussy, T. Fu, J. Guo, S. Bassirian, S. Singh, E. V. Madamba, R. Bettencourt, L. Richards, R. T. Yu,

A. R. Atkins, T. Huan, D. A. Brenner, C. B. Sirlin, M. Downes, R. M. Evans, and R. Loomba. A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis. *Cell Metabolism*, 32(5): 878–888, 2020. doi: 10.1016/j.cmet.2020.06.005.

[172] M. Oudah and A. Henschel. Taxonomy-aware feature engineering for microbiome classification. *BMC Bioinformatics*, 19(1):227, 12 2018. doi: 10.1186/s12859-018-2205-3.

[173] E. Paradis and K. Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35 (3):526–528, 2 2019. doi: 10.1093/bioinformatics/bty633.

[174] D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, and G. W. Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, 2(11):1533–1542, 11 2017. doi: 10.1038/s41564-017-0012-7.

[175] D. H. Parks, M. Chuvochina, P.-A. Chaumeil, C. Rinke, A. J. Mussig, and P. Hugenholtz. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, 38(9): 1079–1086, 9 2020. doi: 10.1038/s41587-020-0501-8.

[176] E. Pasolli, L. Schiffer, P. Manghi, A. Renson, V. Obenchain, D. T. Truong, F. Beghini, F. Malik, M. Ramos, J. B. Dowd, C. Huttenhower, M. Morgan, N. Segata, and L. Waldron. Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, 14(11):1023–1024, 11 2017. doi: 10.1038/nmeth.4468.

[177] E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower, and N. Segata. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*, 176 (3):649–662, 2019. doi: 10.1016/j.cell.2019.01.001.

[178] J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, 12 2013. doi: 10.1038/nmeth.2658.

[179] M. R. Pausan, C. Csorba, G. Singer, H. Till, V. Schöpf, E. Santigli, B. Klug, C. Högenauer, M. Blohs, and C. Moissl-Eichinger. Exploring the Archaeome: Detection of Archaeal Signatures in the Human Body. *Frontiers in Microbiology*, 10, 12 2019. doi: 10.3389/fmicb.2019.02796.

[180] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk. Reducing Misclassification Costs. In *Machine Learning Proceedings 1994*, pages 217–225. Elsevier, 1994. doi: 10.1016/B978-1-55860-335-6.50034-9.

[181] T. L. Pedersen. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks, 2020. URL `https://cran.r-project.org/package=ggraph`.

[182] A. C. Poole, J. K. Goodrich, N. D. Youngblut, G. G. Luque, A. Ruaud, J. L. Sutter, J. L. Waters, Q. Shi, M. El-Hadidi, L. M. Johnson, H. Y. Bar, D. H. Huson, J. G. Booth, and R. E. Ley. Human Salivary Amylase Gene Copy Number Impacts Oral and Gut Microbiomes. *Cell Host and Microbe*, 25(4):553–564, 2019. doi: 10.1016/j.chom.2019.03.001.

[183] B. M. Popkin. The nutrition transition and its health implications in lower-income countries. *Public Health Nutrition*, 1(1):5–21, 1998. doi: 10.1079/phn19980004.

[184] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 3 2010. doi: 10.1038/nature08821.

[185] N. Qin, F. Yang, A. Li, E. Prifti, Y. Chen, L. Shao, J. Guo, E. Le Chatelier, J. Yao, L. Wu, and others. Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516):59–64, 2014.

[186] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, 35(9):833–844, 9 2017. doi: 10.1038/nbt.3935.

[187] S. Rachid Zaim, C. Kenost, J. Berghout, W. Chiu, L. Wilson, H. H. Zhang, and Y. A. Lussier. binomialRF: interpretable combinatoric efficiency of random forests to identify biomarker interactions. *BMC Bioinformatics*, 21(1):374, 12 2020. doi: 10.1186/s12859-020-03718-9.

[188] A. S. Raman, J. L. Gehrig, S. Venkatesh, H.-W. Chang, M. C. Hibberd, S. Subramanian, G. Kang, P. O. Bessong, A. A. Lima, M. N. Kosek, W. A. Petri, D. A. Rodionov, A. A. Arzamasov, S. A. Leyn, A. L. Osterman, S. Huq, I. Mostafa, M. Islam, M. Mahfuz, R. Haque, T. Ahmed, M. J. Barratt, and J. I. Gordon. A sparse covarying unit that describes healthy and impaired human gut microbiota development. *Science*, 365(6449):eaau4735, 7 2019. doi: 10.1126/science.aau4735.

[189] L. J. Revell. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3 (2):217–223, 4 2012. doi: 10.1111/j.2041-210X.2011.00169.x.

[190] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, New York, NY, USA, 8 2016. ACM. ISBN 9781450342322. doi: 10.1145/2939672.2939778.

[191] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-Agnostic Interpretability of Machine Learning. *arXiv*, (Whi), 6 2016.

[192] B. A. Rosa, K. Hallsworth-Pepin, J. Martin, A. Wollam, and M. Mitreva. Genome Sequence of Christensenella minuta DSM

22607T. *Genome Announcements*, 5(2), 1 2017. doi: 10.1128/genomeA.01451-16.

[193] D. Rothschild, O. Weissbrod, E. Barkan, A. Kurilshikov, T. Korem, D. Zeevi, P. Costea, A. Godneva, I. Kalka, N. Bar, S. Shilo, D. Lador, A. Vich Vila, N. Zmora, M. Pevsner-Fischer, D. Israeli, N. Kosower, G. Malka, B. chen Wolf, T. Avnit-Sagi, M. Lotan-Pompan, A. Weinberger, Z. Halpern, S. Carmi, J. Fu, C. Wijmenga, A. Zhernakova, E. Elinav, and E. Segal. Environment dominates over host genetics in shaping human gut microbiota. *Nature Publishing Group*, 555:210–215, 2018. doi: 10.1038/nature25973.

[194] A. Ruaud, S. Esquivel-Elizondo, J. de la Cuesta-Zuluaga, J. L. Waters, L. T. Angenent, N. D. Youngblut, and R. E. Ley. Syntrophy via Interspecies H 2 Transfer between Christensenella and Methanobrevibacter Underlies Their Global Cooccurrence in the Human Gut. *mBio*, 11(1), 2 2020. doi: 10.1128/mBio.03235-19.

[195] J. L. Rychlik and T. May. The Effect of a Methanogen, Methanobrevibacter smithii , on the Growth Rate, Organic Acid Production, and Specific ATP Activity of Three Predominant Ruminal Cellulolytic Bacteria. *Current Microbiology*, 40(3):176–180, 3 2000. doi: 10.1007/s002849910035.

[196] A. B. Sahakian, S.-R. Jee, and M. Pimentel. Methane and the Gastrointestinal Tract. *Digestive Diseases and Sciences*, 55(8): 2135–2143, 8 2010. doi: 10.1007/s10620-009-1012-0.

[197] A. Salonen, J. Nikkilä, J. Jalanka-Tuovinen, O. Immonen, M. Rajilić-Stojanović, R. A. Kekkonen, A. Palva, and W. M. de Vos. Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *Journal of Microbiological Methods*, 81(2):127–134, 5 2010. doi: 10.1016/j.mimet.2010.02.007.

[198] B. S. Samuel, E. E. Hansen, J. K. Manchester, P. M. Coutinho, B. Henrissat, R. Fulton, P. Latreille, K. Kim, R. K. Wilson, and J. I. Gordon. Genomic and metabolic adaptations of Methanobrevibacter smithii to the human gut. *Proceedings of*

the National Academy of Sciences of the United States of America, 104(25):10643–8, 6 2007. doi: 10.1073/pnas.0704189104.

[199] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona. Fiji: an open-source platform for biological-image analysis. *Nature Methods*, 9(7):676–682, 7 2012. doi: 10.1038/nmeth.2019.

[200] M. Schubert. clustermq: Evaluate Function Calls on HPC Schedulers (LSF, SGE, SLURM, PBS/Torque), 2020. URL `https://cran.r-project.org/package=clustermq`.

[201] A. Schwiertz, D. Taras, K. Schäfer, S. Beijer, N. A. Bos, C. Donus, and P. D. Hardt. Microbiota and SCFA in lean and overweight healthy subjects. *Obesity*, 18(1):190–195, 2010. doi: 10.1038/oby.2009.167.

[202] M. R. Segal. Machine Learning Benchmarks and Random Forest Regression. *Biostatistics*, pages 1–14, 2004.

[203] N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9(8):811–814, 8 2012. doi: 10.1038/nmeth.2066.

[204] R. Sender, S. Fuchs, and R. Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, 14(8):e1002533, 8 2016. doi: 10.1371/journal.pbio.1002533.

[205] R. D. Shah and N. Meinshausen. Random intersection trees. *Journal of Machine Learning Research*, 15:629–654, 2014.

[206] B. J. Shapiro, J. B. Leducq, and J. Mallet. What Is Speciation? *PLoS Genetics*, 12(3):1–14, 2016. doi: 10.1371/journal.pgen.1005860.

[207] L. S. Shapley. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press, 12 1953. doi: 10.1515/9781400881970-018.

[208] L. Shen, Q. Zhao, X. Wu, X. Li, Q. Li, and Y. Wang. Interspecies electron transfer in syntrophic methanogenic consortia: From cultures to bioreactors. *Renewable and Sustainable Energy Reviews*, 54:1358–1367, 2015. doi: 10.1016/j.rser.2015.10.102.

[209] M. E. Shiffman, R. M. Soo, P. G. Dennis, M. Morrison, G. W. Tyson, and P. Hugenholtz. Gene and genome-centric analyses of koala and wombat fecal microbiomes point to metabolic specialization for Eucalyptus digestion. *PeerJ*, 5:e4075, 2017. doi: 10.7717/peerj.4075.

[210] C. M. K. Sieber, A. J. Probst, A. Sharrar, B. C. Thomas, M. Hess, S. G. Tringe, and J. F. Banfield. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, 3(7):836–843, 7 2018. doi: 10.1038/s41564-018-0171-1.

[211] A. Spang, E. F. Caceres, and T. J. G. Ettema. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science*, 357(6351):eaaf3883, 8 2017. doi: 10.1126/science.aaf3883.

[212] A. J. M. Stams and C. M. Plugge. Electron transfer in syntrophic communities of anaerobic bacteria and archaea. *Nature Reviews Microbiology*, 7(8):568–577, 8 2009. doi: 10.1038/nrmicro2166.

[213] J. Stewart, V. Chadwick, and A. Murray. Carriage, quantification, and predominance of methanogens and sulfate-reducing bacteria in faecal samples. *Letters in Applied Microbiology*, 43 (1):58–63, 7 2006. doi: 10.1111/j.1472-765X.2006.01906.x.

[214] E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the Performance of Prediction Models. *Epidemiology*, 21(1):128–138, 1 2010. doi: 10.1097/EDE.0b013e3181c30fb2.

[215] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 12 2007. doi: 10.1186/1471-2105-8-25.

[216] Y. Sun, A. K. Wong, and M. S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719, 2009. doi: 10.1142/S0218001409007326.

[217] M. Sundararajan and A. Najmi. The many Shapley values for model explanation. In H. Daumé III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, pages 9269–9278. PMLR, 2020.

[218] S. Suthaharan. Decision Tree Learning. In *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, pages 237–269. Springer US, Boston, MA, 2016. ISBN 978-1-4899-7641-3. doi: 10.1007/978-1-4899-7641-3{_}10.

[219] R. C. Team. R: A Language and Environment for Statistical Computing, 2020. URL `https://www.r-project.org/`.

[220] C. A. Thaiss, S. Itav, D. Rothschild, M. T. Meijer, M. Levy, C. Moresi, L. Dohnalová, S. Braverman, S. Rozin, S. Malitsky, M. Dori-Bachash, Y. Kuperman, I. Biton, A. Gertler, A. Harmelin, H. Shapiro, Z. Halpern, A. Aharoni, E. Segal, and E. Elinav. Persistent microbiome alterations modulate the rate of post-dieting weight regain. *Nature*, 540(7634):544–551, 2016. doi: 10.1038/nature20796.

[221] J. H. Thiele and J. G. Zeikus. Control of Interspecies Electron Flow during Anaerobic Digestion: Significance of Formate Transfer versus Hydrogen Transfer during Syntrophic Methanogenesis in Flocs. *Applied and environmental microbiology*, 54(1):20–29, 1 1988. doi: 10.1128/AEM.54.1.20-29.1988.

[222] A. M. Thomas, P. Manghi, F. Asnicar, E. Pasolli, F. Armanini, M. Zolfo, F. Beghini, S. Manara, N. Karcher, C. Pozzi, S. Gandini, D. Serrano, S. Tarallo, A. Francavilla, G. Gallo, M. Trompetto, G. Ferrero, S. Mizutani, H. Shiroma, S. Shiba, T. Shibata, S. Yachida, T. Yamada, J. Wirbel, P. Schrotz-King, C. M. Ulrich, H. Brenner, M. Arumugam, P. Bork, G. Zeller, F. Cordero, E. Dias-Neto, J. C. Setubal, A. Tett, B. Pardini,

M. Rescigno, L. Waldron, A. Naccarati, and N. Segata. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine*, 25(4):667–678, 4 2019. doi: 10.1038/s41591-019-0405-7.

[223] J. H. Thornton and F. N. Owens. Monensin Supplementation and in vivo Methane Production by Steers. *Journal of Animal Science*, 52(3):628–634, 3 1981. doi: 10.2527/jas1981.523628x.

[224] R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[225] B. D. Topçuoğlu, N. A. Lesniak, M. T. Ruffin, J. Wiens, and P. D. Schloss. A framework for effective application of machine learning to microbiome-based classification problems. *mBio*, 11 (3):1–13, 2020. doi: 10.1128/mBio.00434-20.

[226] W. Tottey, N. Gaci, G. Borrel, M. Alric, P. W. O 'toole, and J.-F. Brug Ere. In-vitro model for studying methanogens in human gut microbiota. *Anaerobe*, 34:50–52, 2015. doi: 10.1016/j.anaerobe.2015.04.009.

[227] S. Traore, S. Khelaifia, N. Armstrong, J. Lagier, and D. Raoult. Isolation and culture of Methanobrevibacter smithii by co-culture with hydrogen-producing bacteria on agar plates. *Clinical Microbiology and Infection*, 25(12):1–1561, 12 2019. doi: 10.1016/j.cmi.2019.04.008.

[228] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457, 2009. doi: 10.1038/nature07540.

[229] W. Turpin, O. Espin-Garcia, W. Xu, M. S. Silverberg, D. Kevans, M. I. Smith, D. S. Guttman, A. Griffiths, R. Panaccione, A. Otley, L. Xu, K. Shestopaloff, G. Moreno-Hagelsieb, A. D. Paterson, and

K. Croitoru. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nature Genetics*, 48(11): 1413–1417, 2016. doi: 10.1038/ng.3693.

[230] B. Upadhyaya, L. McCormack, A. R. Fardin-Kia, R. Juenemann, S. Nichenametla, J. Clapper, B. Specker, and M. Dey. Impact of dietary resistant starch type 4 on human gut microbiota and immunometabolic functions. *Scientific Reports*, 6:1–12, 2016. doi: 10.1038/srep28797.

[231] J. A. van de Pol, N. v. Best, C. A. Mbakwa, C. Thijs, P. H. Savelkoul, I. C. Ilja, M. W. Hornef, M. Mommers, and J. Penders. Gut colonization by methanogenic archaea is associated with organic dairy consumption in children. *Frontiers in Microbiology*, 8(MAR):1–10, 2017. doi: 10.3389/fmicb.2017.00355.

[232] D. Vandeputte, G. Falony, S. Vieira-Silva, R. Y. Tito, M. Joossens, and J. Raes. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut*, 65(1):57–62, 1 2016. doi: 10.1136/gutjnl-2015-309618.

[233] S. Vanderhaeghen, C. Lacroix, and C. Schwab. Methanogen communities in stools of humans of different age and health status and co-occurrence with bacteria. *FEMS Microbiology Letters*, 362 (13):1–8, 2015. doi: 10.1093/femsle/fnv092.

[234] I. Vujkovic-Cvijin, J. Sklar, L. Jiang, L. Natarajan, R. Knight, and Y. Belkaid. Host variables confound gut microbiota studies of human disease. *Nature*, 587(7834):448–454, 11 2020. doi: 10.1038/s41586-020-2881-9.

[235] W.-L. Wang, S.-Y. Xu, Z.-G. Ren, L. Tao, J.-W. Jiang, and S.-S. Zheng. Application of metagenomics in the human gut microbiome. *World Journal of Gastroenterology*, 21(3):803, 2015. doi: 10.3748/wjg.v21.i3.803.

[236] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and

R. Knight. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME Journal*, 10(7): 1669–1681, 2016. doi: 10.1038/ismej.2015.235.

[237] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

[238] P. G. Wolf, G. Parthasarathy, J. Chen, H. M. O'Connor, N. Chia, A. E. Bharucha, and H. R. Gaskins. Assessing the colonic microbiome, hydrogenogenic and hydrogenotrophic genes, transit and breath methane in constipation. *Neurogastroenterology & Motility*, page e13056, 3 2017. doi: 10.1111/nmo.13056.

[239] D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014. doi: 10.1186/gb-2014-15-3-r46.

[240] M. N. Wright and A. Ziegler. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 2017. doi: 10.18637/jss.v077.i01.

[241] M. N. Wright and A. Ziegler. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 2017. doi: 10.18637/jss.v077.i01.

[242] Y.-W. Wu, B. A. Simmons, and S. W. Singer. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607, 2 2016. doi: 10.1093/bioinformatics/btv638.

[243] Y. Yang, N. Chen, and T. Chen. Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model. *Cell Systems*, 4(1):129–137, 2017. doi: 10.1016/j.cels.2016.12.012.

[244] Y. Yang, X. Wang, K. Xie, C. Zhu, N. Chen, and T. Chen. Inferring Multiple Metagenomic Association Networks based on Variation of Environmental Factors. *bioRxiv*, 2020. doi: 10.1101/2020.03.04.976423.

[245] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402): 222–227, 6 2012. doi: 10.1038/nature11053.

[246] N. D. Youngblut, J. de la Cuesta-Zuluaga, and R. E. Ley. Incorporating genome-based phylogeny and trait similarity into diversity assessments helps to resolve a global collection of human gut metagenomes. *bioRxiv*, 2020. doi: 10.1101/2020.07.16.207845.

[247] N. D. Youngblut, G. H. Reischer, S. Dauser, C. Walzer, G. Stalder, A. H. Farnleitner, and R. E. Ley. Strong influence of vertebrate host phylogeny on gut archaeal diversity. *bioRxiv*, 2020. doi: 10.1101/2020.11.10.376293.

[248] G. Zeller, J. Tap, A. Y. Voigt, S. Sunagawa, J. R. Kultima, P. I. Costea, A. Amiot, J. Böhm, F. Brunetti, N. Habermann, R. Hercog, M. Koch, A. Luciani, D. R. Mende, M. A. Schneider, P. Schrotz-King, C. Tournigand, J. Tran Van Nhieu, T. Yamada, J. Zimmermann, V. Benes, M. Kloor, C. M. Ulrich, M. Knebel Doeberitz, I. Sobhani, and P. Bork. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11):766, 2014. doi: 10.15252/msb.20145645.

[249] H. Zhang, J. K. DiBaise, A. Zuccolo, D. Kudrna, M. Braidotti, Y. Yu, P. Parameswaran, M. D. Crowell, R. Wing, B. E. Rittmann, and R. Krajmalnik-Brown. Human gut microbiota in obesity and after gastric bypass. *Proceedings of the National Academy of Sciences of the United States of America*, 106(7): 2365–70, 2 2009. doi: 10.1073/pnas.0812600106.

[250] Y.-H. Zhou and P. Gallins. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. *Frontiers in Genetics*, 10, 6 2019. doi: 10.3389/fgene.2019.00579.

[251] Y.-H. Zhou, K. Xia, and F. A. Wright. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics*, 27(19):2672–2678, 10 2011. doi: 10.1093/bioinformatics/btr449.

[252] Y. Zou, W. Xue, G. Luo, Z. Deng, P. Qin, R. Guo, H. Sun, Y. Xia, S. Liang, Y. Dai, D. Wan, R. Jiang, L. Su, Q. Feng, Z. Jie, T. Guo, Z. Xia, C. Liu, J. Yu, Y. Lin, S. Tang, G. Huo, X. Xu, Y. Hou, X. Liu, J. Wang, H. Yang, K. Kristiansen, J. Li, H. Jia, and L. Xiao. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. *Nature Biotechnology*, 37(2):179–185, 2 2019. doi: 10.1038/s41587-018-0008-8.

[253] Y. Zou, W. Xue, X. Lin, T. Hu, S.-W. Liu, C.-H. Sun, G. Luo, M. Lv, Y. Dai, K. Kristiansen, and L. Xiao. Taxonomic Description and Genome Sequence of Christensenella intestinihominis sp. nov., a Novel Cholesterol-Lowering Bacterium Isolated From Human Gut. *Frontiers in Microbiology*, 12, 2 2021. doi: 10.3389/fmicb.2021.632361.

[254] C. Zuñiga, L. Zaramela, and K. Zengler. Elucidation of complexity and prediction of interactions in microbial communities. *Microbial Biotechnology*, 9 2017. doi: 10.1111/1751-7915.12855.

[255] C. Zuñiga, L. Zaramela, and K. Zengler. Elucidation of complexity and prediction of interactions in microbial communities. *Microbial Biotechnology*, (237897), 2017. doi: 10.1111/1751-7915.12855.

Appendix A
Appendix to Chapters 2 and 3

Relevant author contributions to Ruaud et al. (*in preparation*)

I conceptualized the whole endoR method and received valuable feedback from Dr Niklas Pfister. I implemented the method in R, simulated data to benchmark the method under the supervision of Dr Niklas Pfister, and produced all results. Dr Niklas Pfister and I wrote the mathematical text of the method. I wrote the rest of the manuscript, and Dr Niklas Pfister and Dr Nicholas Youngblut reviewed and edited it. Contributions as defined by CRediT roles (https://casrai.org/credit/) are formally described in Table A.1.

I adapted the current draft of the article into Chapters 2 and 3.

CRediT role[a]	Authors
Conceptualization	Albane Ruaud and Niklas Pfister
Data curation	Albane Ruaud and Nicholas D Youngblut
Formal analysis	Albane Ruaud
Investigation	Albane Ruaud
Methodology	Albane Ruaud and Niklas Pfister
Software	Albane Ruaud
Supervision	Niklas Pfister and Nicholas D Youngblut
Visualization	Albane Ruaud
Writing – original draft	Albane Ruaud and Niklas Pfister
Writing – review & editing	Albane Ruaud, Niklas Pfister and Nicholas D Younglut
Funding acquisition	Ruth E Ley

Table A.1: Author contributions relevant to Chapters 2 and 3.

[a] Contributions are attributed according to CRediT roles (https://casrai.org/credit/).

Additional tables

Method	p	n	B	**ru_utime**	**cpu**	**mem**	**maxvmem**
endoR	18	1000	1	42,50	1737,82	496,74	2,42
SHAP	18	1000	N/A	21,51	31949,77	5274,00	1,15
endoR	18	2000	1	37,57	2786,45	795,01	3,74
SHAP	18	2000	N/A	47,07	159783,36	27706,33	3,48
endoR	18	500	1	38,41	1383,02	363,97	3,95
SHAP	18	500	N/A	20,43	19027,86	2799,00	1,72
endoR	18	1000	1	40,85	1914,13	519,68	3,63
endoR	18	1000	10	73,60	19167,80	5221,00	3,82
endoR	18	1000	20	98,72	34508,98	9327,67	3,66
endoR	50	1000	1	78,80	13336,75	4205,00	3,97
endoR	100	1000	1	144,47	35681,31	12714,33	5,16
SHAP	50	1000	N/A	26,25	54412,93	8517,33	2,57
SHAP	100	1000	N/A	36,63	142459,98	23756,67	3,47
endoR	18	1000	40	179,56	77584,25	21130,00	3,81

Table A.2: Average computation time and memory of endoR and `shap` on artificial phenotypes.

p: number of features; *n*: number of observations; *B*: number of bootstraps (for endoR only, N/A: not applicable; ru_utime: consumed user time, i.e., actual CPU time, in s; cpu: ru_time + system time, in s; mem: total memory · cpu, GB·s; max: maximal memory used at any time, GB. Average values across three replicated runs are reported (each replicate value is plotted on Figure 3.8).

Rule	B	V	s	α	$\hat{y}$	Importance	Multiplicity
s_M_micronuciformis=Low & s_A_equolifaciens=High	99	2	0.228	0	0	0.228±0.040	3.572±0.277
s_M_micronuciformis=High	95	1	0.335	0.07	0.98	0.283±0.069	95.674±50.298
s_L_salivarius=High & s_C_stercoris=Low	93	2	0.265	0	1	0.265±0.043	1.606±0.437
s_Clostridium_sp_ASF356=Low & s_E_ventriosum=Low	93	2	0.207	0	1	0.207±0.034	0.700±0.188
s_K_denitrificans=High & g_Erysipelatoclostridium=High	83	2	0.192	0	1	0.192±0.034	1.001±0.002
s_B_viscericola=Low & s_Clostridium_sp_ASF356=Low	83	2	0.192	0	1	0.192±0.032	0.581±0.214
s_V_parvula=High	81	1	0.340	0.10	0.97	0.263±0.066	40.472±34.339
s_P_enoeca=High & g_Leptotrichia=High	81	2	0.195	0	1	0.195±0.039	1.146±0.391
s_K_denitrificans=High & s_Pantoea_sp_PSNIH2=High	79	2	0.201	0	1	0.201±0.036	0.762±0.509
s_F_caenicola=Medium & s_M_micronuciformis=Low	78	2	0.190	0	0	0.190±0.032	1.338±0.219
s_R_microfusus=Low & f_Pasteurellaceae=High	77	2	0.184	0	1	0.184±0.031	1.496±0.251
g_Campylobacter=High & s_C_stercoris=Low	76	2	0.278	0.07	0.98	0.236±0.062	0.879±0.448
s_C_symbiosum=Medium & s_M_micronuciformis=Low	76	2	0.190	0	0	0.190±0.031	2.600±0.349
g_Leptotrichia=High	75	1	0.342	0.15	0.96	0.230±0.058	18.662±20.274
f_Eggerthellaceae=Low & s_F_nucleatum=High	73	2	0.194	0	1	0.194±0.029	1.124±0.265
s_L_saburreum=Medium & s_M_micronuciformis=Low	73	2	0.193	0	0	0.193±0.030	2.417±0.256
g_Streptococcus=High	70	1	0.344	0.15	0.96	0.235±0.066	25.774±21.531
f_Veillonellaceae=High & g_Erysipelatoclostridium=High	70	2	0.188	0	1	0.188±0.033	1.764±0.442

Table A.3: Stable decision ensemble generated by endoR from the RF predicting cirrhosis vs healthy individuals. B: number of bootstraps in which the decision was selected; V: number of variables in the rule; s: support; $\hat{y}$: prediction (0 = healthy, 1 = cirrhosis). Average ± standard deviation importance and multiplicity.

Variable 1	**Variable 2**	F	Γ
M_micronuciformis__High		10.5924	0.4480
V_parvula__High		3.5552	0.4270
Streptococcus__High		1.7156	0.4067
Leptotrichia__High		1.2759	0.3991
M_micronuciformis__Low		0.4996	-0.2461
L_salivarius__High		0.1709	0.4011
A_equolifaciens__High		0.1464	-0.1802
C_symbiosum__Medium	M_micronuciformis__Low	0.1196	-0.3217
M_micronuciformis__Low	A_equolifaciens__High	0.1135	-0.2175
L_saburreum__Medium	M_micronuciformis__Low	0.1128	-0.3226
K_denitrificans__High		0.0903	0.2615
C_symbiosum__Medium		0.0892	-0.1802
Veillonellaceae__High		0.0843	0.2538
L_saburreum__Medium		0.0839	-0.1795
L_salivarius__High	C_stercoris__Low	0.0810	0.4060
Campylobacter__High		0.0643	0.3695
Pasteurellaceae__High		0.0629	0.2284
F_caenicola__Medium	M_micronuciformis__Low	0.0596	-0.3108
Erysipelatoclostridium__High		0.0551	0.1053
C_stercoris__Low		0.0527	0.0907
R_microfusus__Low	Pasteurellaceae__High	0.0508	0.2610
C_sp_ASF356__Low		0.0492	0.1919
Veillonellaceae__High	Erysipelatoclostridium__High	0.0462	0.2293
F_caenicola__Medium		0.0458	-0.1800
R_microfusus__Low		0.0441	0.1603
F_nucleatum__High		0.0385	0.1770
K_denitrificans__High	Erysipelatoclostridium__High	0.0351	0.2690
E_ventriosum__Low		0.0317	0.2182
K_denitrificans__High	P_sp_PSNIH2__High	0.0292	0.2881
P_enoeca__High	Leptotrichia__High	0.0282	0.2101
C_sp_ASF356__Low	E_ventriosum__Low	0.0274	0.2773
Eggerthellaceae__Low	F_nucleatum__High	0.0269	0.1857
Campylobacter__High	C_stercoris__Low	0.0254	0.3713
B_viscericola__Low		0.0243	0.2180
B_viscericola__Low	C_sp_ASF356__Low	0.0236	0.2983
P_sp_PSNIH2__High		0.0217	0.1417
Eggerthellaceae__Low		0.0211	0.0968
P_enoeca__High		0.0170	0.0760

Table A.4: Variable and interaction importance and influence calculated by endoR from the RF predicting cirrhosis vs healthy individuals. F: importance; Γ: influence.

Additional figures

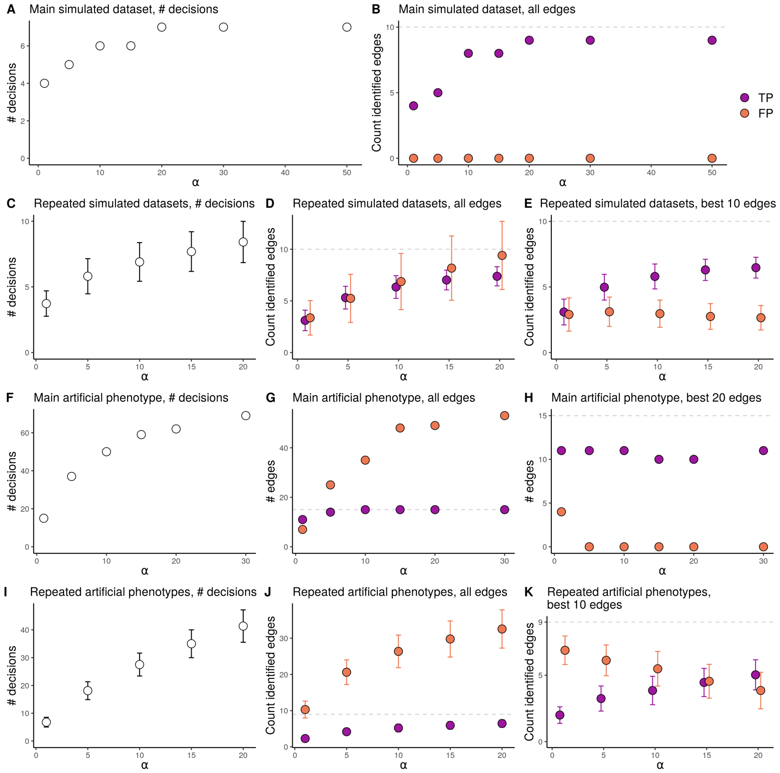

Figure A.1: Effect of α on the number of decisions, TP and FP in the stable decision ensemble. A-E/ Simulated data with $n = 1000$, $r = 0.05$, and $\pi = 0.7$; A-B/ single replicate with $B = 100$ (Fig 3.2), C-E/ 100 repetitions with $B = 10$ (Fig 3.6). F-K/ Artificial phenotypes with $r = 0.05$; F-H/ single replicate with $B = 100$ (Fig 3.3); I-K/ repetitions of artificial phenotypes with $B = 10$ (Fig 3.6). The means (points) and standard deviations (bars) are plotted for repetitions on C-E and I-K. The dotted grey lines correspond to the total number of true edges.

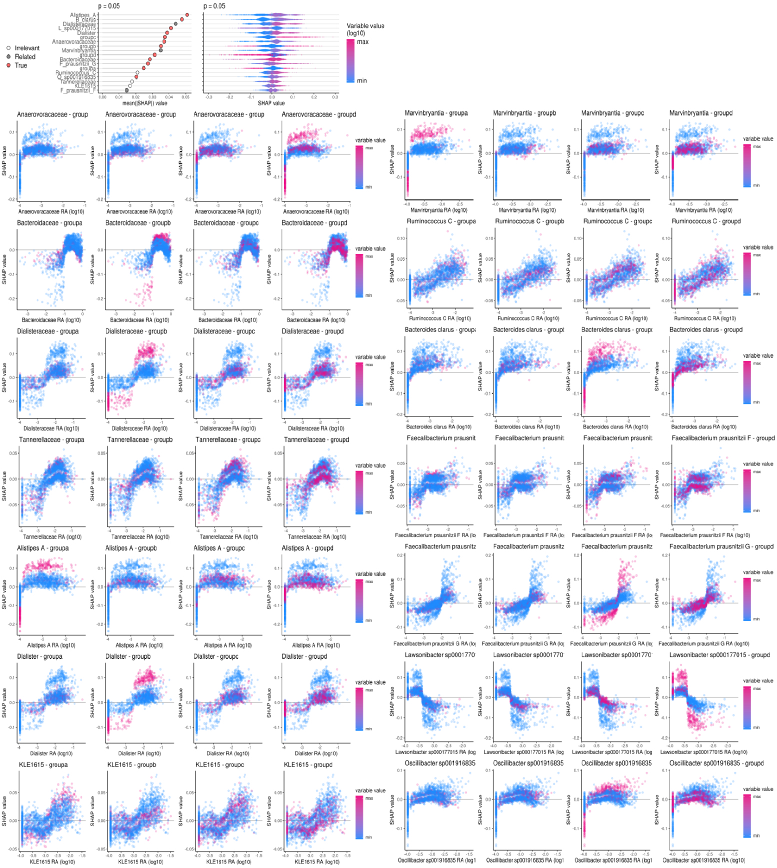

Figure A.2: SHAP values from a random forest classifier fitted on metagenomes with an artificial phenotype. SHAP values were calculated from the random forest classifier fitted on the replicate of metagenomes with an artificial phenotype presented in Figure 3.3. I used the iBreakDown R-package [18] to calculate SHAP values. The feature importance corresponds to the average of the absolute SHAP values across samples. As the SHAP interaction values could not be calculated, SHAP values for the relative abundances (log10 transformed) of all 14 taxa colored group categories are displayed.

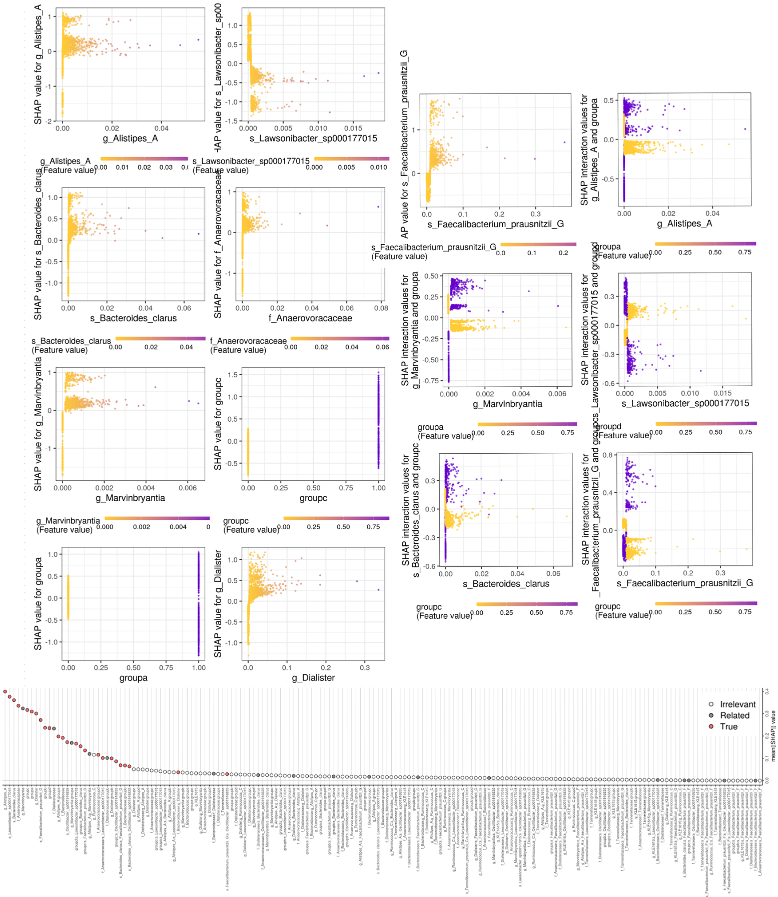

Figure A.3: SHAP values from the XGBoost classifier fitted on metagenomes with an artificial phenotype. SHAP values were calculated from the XGBoost classifier trained on the replicate of metagenomes with an artificial phenotype presented in Figure 3.3. The feature importance is given by the average of the absolute SHAP values across samples. Given the high number of features and interactions, we only plotted the first 10 individuals and 4 interactions, as ranked by feature importance; those plots are the direct outputs from the SHAPforxgboost R package.

Appendix B
Appendix to Chapter 4

Additional tables and figures

Metadata	n^{a}	Summary[b]
Dataset	2203	See Table B.2
Number of reads	2203	1.42 - 356.01 million reads (52.18 and 58.92±43.69 million reads)
Country	2203	See Table B.2
Region[c]	2203	Africa (190), Central Asia (377), Europe (1159), Fiji (115), India (88), North America (219), Peru (55)
Westernized	2203	False (517), True (1686)
Gender	1780	Female (1109), Male (671)
Age	1675	19 - 84 years old (33 and 40.34±17.70 years old)
BMI	1020	16.02 - 36.41 kg.m^{-2} (23.27 and 24.03±3.76 kg.m^{-2})
Enterotype	2203	ETF (882), ETB (622), ETP (699)

Table B.1: Metadata included in predictive models training

[a] Number of samples with available information.

[b] For numeric variables: minimal - maximal values (median and mean ± standard deviation). For categorical variables: each level (number of samples in the level).

[c] Samples from countries of a same geographic area. Region with a unique country are named after the country name to prevent confusion.

Dataset	Country[a]	n	n metadata[b]
AsnicarF, 2017	ITA	8	0
Bengtsson-PalmeJ, 2015	SWE	70	0
BritoIL, 2016	FJI	115	0
CosteaPI, 2017	DEU	2	0
CosteaPI, 2017	KAZ	21	21
DhakanDB, 2019	IND	88	88
FengQ, 2015	AUT	12	12
HanniganGD, 2017	CAN	3	0
HanniganGD, 2017	USA	24	0
HansenLBS, 2018	DNK	204	25
Heitz-BuschartA, 2016	LUX	2	2
HMP, 2012	USA	137	0
JieZ, 2017	CHN	107	100
KarlssonFH, 2013	DEU	2	0
KarlssonFH, 2013	FRA	1	0
KarlssonFH, 2013	ISL	1	0
KarlssonFH, 2013	SWE	39	0
LiJ, 2017	CHN	41	0
LiuW, 2016	MNG	110	0
LouisS, 2016	DEU	92	0
Obregon-TitoAJ, 2015	PER	7	5
Obregon-TitoAJ, 2015	USA	19	19
PasolliE, 2018	MDG	107	93
PehrssonE, 2016	PER	48	0
PehrssonE, 2016	SLV	71	0
RaymondF, 2016	CAN	36	36
SchirmerM, 2016	NLD	405	396
TettAJ, 2019, a	TZA	36	0
TettAJ, 2019, b	GHA	23	0
TettAJ, 2019, c	ETH	24	0
XieH, 2016	GBR	250	0
YeZ, 2018	CHN	45	45
YuJ, 2015	CHN	53	0

Table B.2: Datasets and country of origins of samples used for analysis

[a] Countries are designated by their ISO 3166 alpha-3 three-letter country code.

[b] Number of samples for which both age and BMI were reported.

Samples[a]	**Model**[b]	**Model parameter**	γ	**k**	**Accuracy**[c]	**Kappa**[c]	p **FS**[c]
all	RF cs	250 trees	0.5	1	0.8270±0.0084	0.6052±0.0190	298.6±4.55
all	RF cs	500 trees	0.8	1	0.8268±0.0085	0.6049±0.0208	287.2±8.56
all	RF	500 trees	0.5	1	0.8368±0.0120	0.6003±0.0315	298.6±4.55
all	RF	1000 trees	0.9	1	0.8362±0.0126	0.5984±0.0332	286.6±6.11
all	RF	250 trees	0.2	1	0.8347±0.0134	0.5960±0.0330	344.5±7.79
all	XGBoost	100 rounds	0.1	0.4	0.8067±0.0164	0.5449±0.0432	NA
all	XGBoost	1000 rounds	0.1	0.4	0.8041±0.0201	0.5393±0.0506	NA
sub	RF cs	500 trees	0.2	0.8	0.8036±0.0265	0.5569±0.0546	112.7±4.60

Table B.3: Predictive performance of models trained to predict the presence of *Methanobacteriaceae* from metagenomes, averaged across 10x cross-validation (CV) 70-30 % train-test sets.

[a] Models trained on CV sets from all samples (n = 2203 samples in total, train = 1542 and test = 661 samples), or only the set of samples with complete information for age, gender and BMI (n = 748 samples in total, train = 524 and test = 224 samples).

[b] Random forest (RF) were fitted using the ranger R-package [240], with sample weights provided to increase the probability of samples from the under-represented class to be sampled at each bootstrap (RF), or with class weights provided to penalize wrong predictions on the under-represented class (RF cs). Gradient boosted model (XGBoost) were fitted using the XGBoost R-package [35].

[c] Number of selected features: average±standard deviation across the ten CV sets.

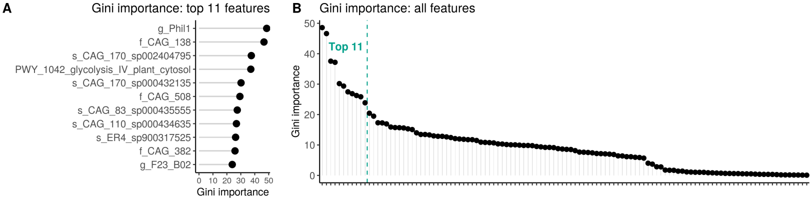

Figure B.1: Gini importances of features in the model described in section 4.3.3.

Appendix C
Appendix to Chapter 5

Author contributions

Hereafter are the relevant author contributions to Ruaud and Esquivel-Elizondo, 2020 [194], from which Chapter 5 is derived. Dr Sofia Esquivel-Elizondo and I performed all *in vitro* experiments and wrote the text for those experiments; I performed all statistical analyses and wrote the text for them; Pr Ruth E Ley, Dr Sofia Esquivel-Elizondo and I wrote the introduction and discussion of the manuscript; Ruth E Ley provided much feedback on the results and methods paragraphs; Nicholas D Youngblut, Jillian L Waters and Lars T Angenent provided feedback on the manuscript. Contributions as defined by CRediT roles (https://casrai.org/credit/) are formally described in Table C.1. I adapted the original published manuscript [194] into Chapter 5.

CRediT role	Authors
Conceptualization	Albane Ruaud, Sofia Esquivel-Elizondo, Ruth E Ley and Lars T Angenent
Investigation	Albane Ruaud and Sofia Esquivel-Elizondo
Formal analysis	Albane Ruaud and Sofia Esquivel-Elizondo
Supervision	Sofia Esquivel-Elizondo and Ruth E Ley
Data curation	Albane Ruaud
Visualization	Albane Ruaud
Writing – original draft	Albane Ruaud, Sofia Esquivel-Elizondo, Ruth E Ley
Writing – review & editing	Albane Ruaud, Sofia Esquivel-Elizondo, Ruth E Ley, Nicholas D Younglut, Lars T Angenent and Jillian L Waters
Funding acquisition	Ruth E Ley and Lars T Angenent

Table C.1: Author contributions relevant to Chapter 5. Contributions are attributed according to CRediT roles (https://casrai.org/credit/).

Additional figures

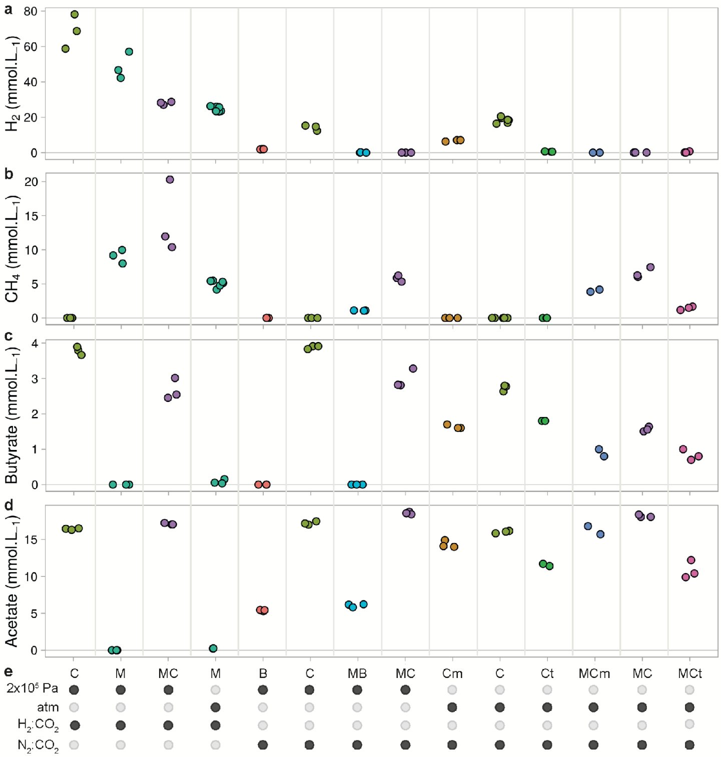

Figure C.1: Summary of gases and SCFA produced in mono- and co-cultures of *C. minuta*, *C. timonensis*, *C. massiliensis*, *B. thetaiotaomicron*, and *M. smithii* after 6 days of growth. a-d/ Points: concentration of each biological replicate after 6 days of growth in all mono- and co-cultures presented in this study (batches 1-4, Table 5.1). e/ Summary of the culture conditions: gas mixture (H_2:CO_2 or N_2:CO_2 80:20 % v/v), initial pressure (2 bar or atmospheric) and microorganisms inoculated. C: *C. minuta*. Ct: *C. timonensis*. Cm: *C. massiliensis*. B: *B. thetaiotaomicron*. M: *M. smithii*. Samples inoculated with the same microorganisms are the same color.

Additional methods

Data and code availability

The jupyter notebooks and associated data are available on GitHub at: https://github.com/aruaud/Ruaud_EsquivelElizondo.

Confocal imaging, equipment, and settings

For confocal microscopy, SYBR® Green I staining was performed as previously described [125] with the following modifications: 0.5 mL of culture were sampled and pelleted by centrifugation for 6 min at 6,000 xg (Benchtop centrifuge, Eppendorf, Hamburg, Germany) and pellets were resuspended in a solution containing 744 μL 1x PBS, 16 μL 25x SYBR®®Green I (Sigma-Aldrich, Merck, Germany) and 40 μL 70 % v/v ethanol. Samples were pelleted and resuspended before imagining in 100 μL 1x PBS, of which 5 μL were immobilized on 50 μL solid agar (1.5 % noble agar in distilled water) [75]. Imaging was performed with a confocal microscope (LSM 780 NLO, Zeiss) using oil and water objectives (40x and 63x). A DPSS laser at 405 nm was used to excite the F420 enzyme of *M. smithii.* Autofluorescence emission was collected on a 32 channel GaAsP array from 455 to 499 nm. A transmitted light detector (T-PMT) was used to collect the whole light spectrum to create a bright field image. On a second track, an Argon laser at 488 nm was used to excite SYBR® Green I and its emission was collected from 508 to 588 nm with the 32 channel GaAsP array as well.

Images were acquired with a time and space resolution of 2048x2048x(1 to 12)x (xyzt) and pixel dimensions of 0.1038x0.1038 μm for the images taken with the x40 oil objective and pixel dimensions of 0.0659x0.0659 μm for the images taken with the x63 oil objective. The bit depth was 16-bit. Acquisition was performed at 20 °C.

Processing of the confocal images

FIJI [199] was used to process the confocal micrographs. Contrast and brightness adjustment were applied to the whole image. Due to the thickness of the aggregates of *Christensenella minuta*, the SYBR® Green I fluorescence intensity was varied with different focal planes. We used a gamma transformation (with gamma = 0.50) to homogenize the fluorescence intensity. The exact same transformation was applied to all samples, even though there were no aggregates, for consistency purposes. Similarly, we applied a gamma transformation to the F420 autofluorescent channel to decrease the low fluorescence coming from SYBR® Green I (gamma = 1.20 to 1.50). As their excitation and emission spectra overlap, there wasa low fluorescence intensity of the SYBR® Green I on the F420 autofluorescent channel. The lookup tables (LUT) were Cyan Hot for the F420 autofluorescence and red (linear LUT, covering the full range of the data) for the SYBR® Green I fluorescence.

Preparation of samples for scanning electron microscopy

Pellets were washed 3-5 times with 1x PBS and then fixed with a 2.5 % v/v glutaraldehyde solution in 1x PBS for 1-2 h at room temperature and post-fixed with 1 % w/v osmium tetroxide for 1 h on ice. Samples were dehydrated in a graded ethanol series followed by drying with CO_2in a Polaron critical point dryer (Quorum Technologies, East Sussex, UK). Finally, cells were sputter coated with a 5 nm thick layer of platinum (CCU-010 Compact coating unit, Safematic GmbH, Bad Ragaz, SWI).

Screening of the short and medium chain fatty acids produced

Before carrying out the experiments presented in the main text, we used gas chromatography (GC) to determine which fatty acids were produced by the cultures and if the corresponding peaks were present in growth medium (brain heart infusion medium). For this screening, the external standards included equimolar mixtures of acetate, propionate, iso-butyrate, butyrate, iso-valerate, valerate, iso-caproate, caproate, heptanoate, and caprylate, from 0.2 to 7 mM. Measurements were performed with a 7890B GC system (Agilent Technologies Inc., Santa Clara, USA) equipped with a capillary column (DB-Fatwax UI 30 m x 0.25 m; Agilent Technologies) and an flame ionization detector detector with a ramp temperature program (initial temperature of 80 °C for 0.5 min, then increasing by 20 °C per minute up to 180 °C, and final temperature of 180 °C for 1 min). The injection and detector temperatures were 250 and 275 °C, respectively. Samples were prepared as for high performance liquid chromatography (HPLC, Methods in the main text) with the addition of an internal standard (Ethyl-butyric acid) and acidification down to pH 2 with 50 % formic acid. Data were acquired and analyzed with the Agilent OpenLAB CDS software.

Only acetate and butyrate were detectedin the mono- and cocultures, and none of the other short and medium chain fatty acids used as standards were detected. As formate was used to acidify samples for the GC measurements, to assess if it was a main product in the cultures, its concentration was measured by HPLC. We also looked for ethanol using HPLC but similar to formate, it was not detected in any of the cultures. Thus, for the experiments in the main text, only acetate and butyrate were quantified via HPLC.

BHI medium showed peaks corresponding to 0.33 mM formate and 6 mM of acetate, which were subtracted from the reported concentrations of the cultures.